DOI:
10.1016/j.websem.2015.04.001

Document Version
Peer reviewed version

Accepted Manuscript

The rationale of PROV

Luc Moreau, Paul Groth, James Cheney, Timothy Lebo, Simon Miles

PII: S1570-8268(15)00017-7
DOI: http://dx.doi.org/10.1016/j.websem.2015.04.001
Reference: WEBSEM 367

To appear in: *Web Semantics: Science, Services and Agents on the World Wide Web*

Received date: 28 February 2014
Revised date: 19 March 2015
Accepted date: 3 April 2015

Please cite this article as: L. Moreau, P. Groth, J. Cheney, T. Lebo, S. Miles, The rationale of PROV, *Web Semantics: Science, Services and Agents on the World Wide Web* (2015), http://dx.doi.org/10.1016/j.websem.2015.04.001

The Rationale of PROV☆

Luc Moreau[a], Paul Groth[b], James Cheney[c], Timothy Lebo[d], Simon Miles[e]

[a]*Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ UK*
[b]*Elsevier Labs, Amsterdam, NL*
[c]*Laboratory for Foundations of Computer Science, University of Edinburgh, Edinburgh EH8 9AB UK*
[d]*Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA*
[e]*Department of Informatics, King's College London, Strand, London WCR2 2LS, UK*

Abstract

The family of documents are the final output of the World Wide Web Consortium Provenance Working Group, chartered to specify a representation of provenance to facilitate its exchange over the Web. This article reflects upon the key requirements, guiding principles, and design decisions that influenced the family of documents. A broad range of requirements were found, relating to the key concepts necessary for describing provenance, such as resources, activities, agents and events, and to balancing 's ease of use with the facility to check its validity. By this retrospective requirement analysis, the article aims to provide some insights into how turned out as it did and why. Benefits of this insight include better inter-operability, a roadmap for alternate investigations and improvements, and solid foundations for future standardization activities.

Keywords: provenance, , standardization, requirement, design decision, rationale

1. Introduction

> *"Provenance is a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing. In particular, the provenance of information is crucial in deciding whether information is to be trusted, how it should be integrated with other diverse information sources, and how to give credit to its originators when reusing it. In an open and inclusive environment such as the Web, where users find information that is often contradictory or questionable, provenance can help those users to make trust judgements. [1]"*

The concept of provenance has been investigated under various names by various computer science communities since the eighties [2, 3, 4, 5]. A recent focus of research on provenance has been its representation and sharing, so as to explain the origin of resources on the Web. This resulted in *adhoc community events* to understand the essence of provenance [6, 7] and define a provenance data model [8]. They were followed by more structured activities such as the World Wide Web Consortium (W3C) Provenance Incubator [9], which paved the way to a standardization effort by the W3C Provenance Working Group. The final output of this formal process resulted in , a data model for provenance on the Web, described by a family of 13 documents, including an overview [10], a primer [11], four Recommendations [1] [12] [13] [14], six technical notes [15] [16] [17] [18] [19] [20], and an implementation report [21].

Whereas the W3C Recommendations and Notes focus on the technical specification of , and publications such as [22] focus on the use and practical deployment of , this article, in contrast, is concerned with the rationale for . This article continues a tradition of similar rationale papers for Semantic Web standardization activities (see [23] for OWL and see [24] for SKOS). It builds on the answers the authors wrote up in response to public reviews during the standardization activity.

Unlike other standardization activities (such as OWL and SKOS), the Provenance Working Group was not chartered to elicit scenarios and requirements, since this task had previously been undertaken by the W3C Provenance Incubator grop [9, 25]. However, through its 8820 public emails[1], 666 issues[2], 600 wiki pages[3], 6000 mercurial commits[4], and 152 teleconferences[5], the Provenance Working Group had numerous rich discussions, adopted guiding principles, considered alternative designs, referred to implicit requirements, and ultimately made design decisions, which help explain why turned out to be as it is. The purpose of this article is to provide justifications for the design of and link it to explicit requirements.

☆This document's provenance can be found at http://eprints.soton.ac.uk/375233/7/provenance.ttl using <http://openprovenance.org/documents#20892220-a071-4ef3-a799-3056447ec8a2> as prov:has_anchor.

[1]Email archive: http://lists.w3.org/Archives/Public/public-prov-wg/
[2]Tracker: http://www.w3.org/2011/prov/track/
[3]Wiki: http://www.w3.org/2011/prov/wiki/Main_Page
[4]Mercurial: https://dvcs.w3.org/hg/prov/
[5]Teleconferences: http://www.w3.org/2011/prov/wiki/Meetings and http://www.w3.org/2011/prov/wiki/PIL_OWL_Ontology#Meeting_notes

We believe that making such requirements explicit is important. Indeed, a benefit for users of is that the model is more likely to be used consistently, if there is a canonical rationale explaining the intentions behind the concepts. This in turn means that should be more interoperable.

For the research community, this article helps position future novel work since the article identifies gaps and aspects that have explicitly been ruled out or considered out of scope for a standardization activity. It also makes it easier to present alternative designs addressing specific existing requirements.

Finally, future standardization processes can build on an explicit presentation of the rationale: charters can list these to scope future activities, and future working groups can further refine requirements, to justify their own work.

1.1. Naming Convention

Terminology evolved during the lifetimes of the W3C Provenance Incubator and Working groups. In this article, we adopt the terminology defined in the W3C Recommendations for to avoid confusion. Thus, requirements that pre-date the standard definitions have been rewritten, to adopt a form that is consistent with the Recommendations.

Likewise, the name was adopted some six months into the lifetime of the standardization activity (see R-2011-09-15/2[6]). Again, for clarity, we use it consistently here in the formulation of all requirements.

A couple of name changes are worth noting: The term "process execution" is now referred to as " activity", whereas "artifact" is now referred to as " entity". Likewise, "recipe" is now called " plan".

1.2. Article Outline

The rest of this article is organized as follows. In Section 2, we summarize the key concepts of that are needed for this article, and we provide a small example to illustrate the data model. In Section 3, we discuss various initiatives related to provenance that precede the creation of the Provenance Working Group. These initiatives are important because they resulted in some deep understanding of provenance issues, and help build a community of expertise and momentum, necessary for the standardization activity. Section 4 focuses on the first provenance-related activity taking place under the auspice of the World Wide Web Consortium: the W3C Provenance Incubator was instrumental in recommending the launch of a standardization activity. Section 5 introduces a categorization of requirements. The Incubator Group drafted a charter, which essentially forms a set of initial requirements for : these are presented in Section 6. Then, Section 7 contains the bulk of this article's contribution: the retrospective requirement analysis of . Finally, in Section 8, we look at aspects that potential future standardization activities may focus on, before concluding the article.

[6]Resolution 2011-09-15/2: http://www.w3.org/2011/prov/meeting/2011-09-15#resolution_2

2. PROV Overview

The family of documents is a set of specifications allowing provenance to be modelled, serialised, exchanged, accessed, merged, translated, and reasoned over. This set includes a conceptual data model [1], an OWL ontology [14], XML serialization [15], a human-readable notation [12], a formal semantics of the conceptual model [17], a set of constraints and inference rules [13], and a mapping to Dublin Core [16]. In this section, we give a brief intuition of the key concepts in the conceptual model using an example.

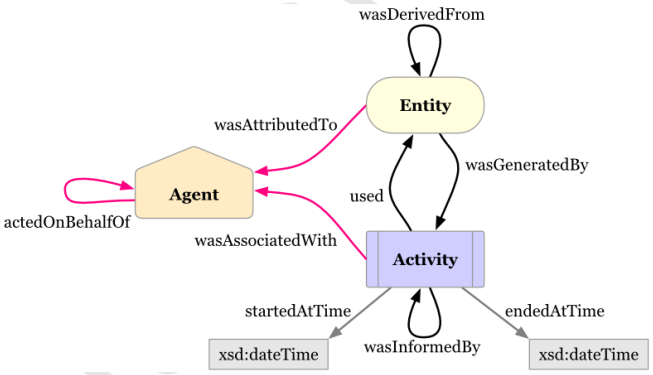

Figure 1: The core concepts of (taken from [14])

Figure 1 shows the core concepts of the data model, centered around the notions of *entity*, a digital, physical or other thing; *activity*, an action using or creating entities; and *agent*, something responsible for an activity taking place as it did.

Consider a scenario, variant of the primer [11], in which an online newspaper publishes an article with a chart about crime statistics based on a data set published by a government. As shown in Figure 2, the article, the chart and the data set are all entities. The process of compiling the chart from the data set is an activity; we say that this activity *used* the data set, that the chart *was generated by* the activity, and that the chart *was derived from* the data set. further allows us to express that the compilation activity *started at* and *ended at* specified times. The compilation activity followed on from another previous activity, the publishing of the data set, and so we may say that the compilation activity *was informed by* the publication activity. The publishing of the data set was the responsibility of a person (agent) called Edith, and we express this by saying that the activity *was associated with* Edith. Edith did not do this independently, but rather *acted on behalf of* the government in publishing the data set. Finally, we can draw a direct connection between Edith and the data set by saying that the data set *was attributed to* Edith, meaning that she was responsible for its creation. As implied by the form of Figure 1, data can be visualized as a directed labelled graph in which nodes are entities, activities and agents and edges represent influences between each of these due to past events (plus annotations of nodes and edges, such as timestamps). A graph visualization for the example above is shown in Figure 2, where entities are shown as ovals, activities are rectangles, and agents are pentagons.

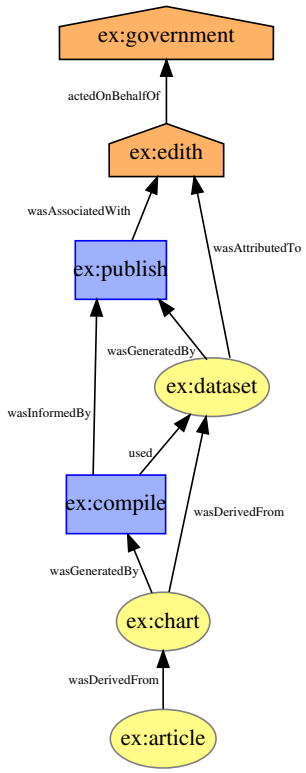

Figure 2: Example PROV graph (turtle file `example.ttl` available from submission)

3. Pre-Standardization Initiatives

In this section, we summarize initiatives that precede the activities that took place at the World Wide Web Consortium. These initiatives include work on provenance in the database, workflow, and Semantic Web communities, and the Provenance Challenge series.

3.1. Database Provenance

Concepts such as source tracking, lineage, and provenance were investigated in databases as early as 1990 [26], and have been studied more intensively over the past 15 years, due in part to the increasing importance of databases in scientific settings, such as bioinformatics [27]. Broadly, database research concerning provenance has focused on three high-level questions:

- How to define and manage provenance information for explaining database query results. Most work in this area proposes an alternative query semantics in which values or records are tagged with additional annotations that are propagated through the query, leading to notions such as Wang and Madnick's Polygen model [26], Cui et al.'s lineage [28], Buneman et al.'s why- and where-provenance [29], and Green et al.'s how-provenance [30]. By placing distinct annotations on the input, the annotations propagated to the result can also be viewed as associating parts of the output with parts of the input. For example, where-provenance annotations are essentially links to the sources of copied data in the input, whereas lineage and why-provenance are tuple-level annotations that indicate sets of input records that suffice to "justify" a record's presence in the output, and how-provenance provides a finer-grained explanation showing how an output tuple was produced by relational projection, selection or join operations on input relations. See [4] for a survey of this area, and Geerts et al. [31] for an adaptation of these ideas to SPARQL.

- How to model and manage provenance for databases as they evolve over time. This area has received less attention than the other two; some contributions include work on tracking where-provenance for manually curated databases [32] and data archiving and versioning [33]. Buneman et al. give an overview of the issues of provenance for evolving data [27].

- How to manage and query provenance information obtained from other systems (e.g. workflow provenance, OPM or) within a database. See e.g. [2, 3] for surveys of this area.

3.2. Workflow Provenance

The development of workflow engines, particularly when applied to enacting reproducible scientific experiments using online data, has been a strong driver in the development of provenance models [34]. A workflow comprises services (or functions, databases, tools, libraries etc.) linked together into a process defined in a user-accessible form, in which the user does not need to be concerned with details of computation such as asynchronous communication, data format conversion, data staging or scheduling. A workflow engine is the software framework which enacts the workflow process, calling each service and passing data between them.

There are several reasons why workflows and provenance are so tightly related. First, reproducibility is a key aim of scientific experiments, and so a record must be kept of what occurs during enactment [35]. Second, as the workflow is created by the end user, they are aware of its structure, and so a record of the process enacted can be readily understood by them and be helpful in interpreting the workflow results. Third, there is a central component, the workflow engine, which can be easily instrumented to include automatic provenance capture. Finally, workflows are often distributed, as the engine makes calls to remote services as part of the process, so interoperable cross-service records are required and local logging is inadequate.

The key concepts in workflow provenance are those of the steps of the workflow, and the inputs and outputs from each step. Provenance data that documents workflow enactment typically describes a directed graph, with steps and data as nodes and input and output connections as edges. Under various forms and with various features, this general model has been used in many workflow engines including REDUX [36], ZOOM [37], Karma [38], Kepler [39], WINGS [40] and Taverna [41], and ontologies for describing workflows such as WDO-It! [42]. A connected strand of work, using a similar general model, considers the provenance of workflows themselves, such as how

they are modified over time by users, as in VisTrails [43], or how they are transformed from an abstract to an executable form, such as has been applied to Pegasus [44].

Therefore, the key influence of workflow provenance efforts on 's development was to include concepts of activities producing and consuming data. In addition, as the data processed in scientific workflows is often in the form of large data sets, the modelling of collections and their elements was also considered important. Despite the influence, is not specific to workflow provenance, nor does it attempt to model all workflow-specific concepts, such as a workflow tasks, ports, and channels. Recently, was extended with new constructs to model such workflow structures [45]. Also note that provenance as modelled in is a generic concept, and can represent activities of humans as readily as software processes enacted by workflow engines.

3.3. The Provenance Challenge Series

During a discussion on provenance standardization at the International Provenance and Annotation Workshop (IPAW'06) [46], participants agreed they needed to understand the different representations used for provenance, its common aspects, and the reasons for its differences. As a result, a "Provenance Challenge" was set to compare and understand existing approaches.

The first provenance challenge [47] was published in June 2006 and concluded in a workshop held in September 2006 in Washington, DC. A simple workflow [6], inspired from a Functional Magnetic Resonance Imaging experiment, formed the basis of the challenge. The workflow consisted of a number of steps, each taking some data as input and producing other data as output. The workflow was not defined in terms of any particular technology such as workflow or programming language. Instead, participants were free to apply their technology of choice. Participants were tasked to contribute: *(i)* a representation of the workflow in their system; *(ii)* a representation of the provenance produced when running the workflow; *(iii)* a representation of the result obtained when running a set of identified provenance queries.

A total of 17 teams [6] contributed a diverse range of results. They decided to hold a second challenge, for which the focus would be interoperability between systems. The first provenance challenge workflow became a de facto benchmark for the provenance community.

The second provenance challenge [48] commenced on December 2006 and concluded in June 2007 with a workshop at High Performance Distributed Computing in Monterey, California, where teams presented and discussed the results. In the second challenge, it was assumed that, within the same workflow, steps were executed by different systems. Teams were tasked to share provenance data produced by their own system, and to perform queries over compositions of provenance data from other teams, as if it had been produced by their own system. The goal was very ambitious and taken up by 14 teams. The second provenance challenge concluded with discussions, out of which a consensus about a common data model began to emerge. This consensus, summarised in the workshop minutes [49], has led to a proposed specification of a provenance data model and inference rules, the Open Provenance Model: OPM. Outside a formal standardization body, the community organized reviews, and revisions of the document, which ultimately led to its publication [8]. OPM was the first community-driven model for provenance. It was itself the focus of the third provenance challenge.

The third provenance challenge [50] was launched in March 2009 to evaluate OPM practically, from an interoperability viewpoint. It resulted in a workshop to discuss findings in June 2009 [7]. Systems were able to export OPM-based provenance, exchange it, and import provenance generated by others. It demonstrated that provenance inter-operability, as envisioned by the Provenance Challenge, was achievable and thus mature enough to begin standardization by an organization like W3C.

3.4. Ontologies for Provenance

Within the Semantic Web community, several ontologies for provenance were produced before the W3C standardization effort. Many of these ontologies fed into the Provenance Incubator Group defining the need for a shared representation. Here, we discuss these ontologies, highlighting their relationship to .

The Proof Markup Language (PML) is an interlingua, grounded in proof theory, designed for the sharing of explanations within knowledge based systems [51]. While originally focused on these applications, PML was latter modularized and expanded to deal with applications from the science and intelligence communities [52]. A revised version of PML, PML3, is being developed which extends [7].

Provenir is a provenance ontology designed to address the needs of e-Science applications [53]. Like PML, it adopts a modular approach. It specifically relies on the philosophical notion of occurrent and continuant, and a similar distinction arises within (with Activity and Entity, respectively). Another ontology that supports provenance within e-Science applications is the SWAN biomedical discourse ontology [54], with a particular focus on the authorship and attribution lifecycle. The provenance portion of SWAN has been separated into the Provenance, Authoring and Versioning (PAV) ontology, which extends to offer specific attribution definitions [55]. Sahoo provides an overview of specific biomedicine ontologies and their usefulness for provenance [56].

Within the library and archival community, provenance has been of longstanding concern [57]. Hence, there are a number of ontologies related to provenance or featuring provenance concepts stemming from that community. The PREservation Metadata: Implementation Strategies (PREMIS) data dictionary is focused on the preservation aspects of digital objects[8]. Dublin Core Metadata Terms[9] is probably the most widely used vocabulary that contains provenance concepts. However, because it is a generic metadata vocabulary, it does not cater for

[7] PML3.0: http://inference-web.org/wiki/PML_3.0
[8] PREMIS: http://www.loc.gov/standards/premis/
[9] DCMI Terms: http://dublincore.org/documents/dcmi-terms/

the expression of some provenance concepts. The Provenance Working Group cooperated with the Dublin Core Metadata Initiative to define a mapping between and Dublin Core [16], and this mapping has since become a DCMI recommended resource[10].

There are a number of ontologies that have been specifically developed to support provenance within Linked Data. This includes the Provenance Vocabulary [58], the Changeset Vocabulary[11], and an OWL version of OPM — OPMV [59]. The Provenance Vocabulary has been refactored to extend specifying classes and properties related to manipulating data items derived from Web resources. While not specifically designed for provenance, the Vocabulary of Interlinked Datasets (VoID) [60] is important to note in this context as it provides a widely used container for metadata. Provenance vocabularies are often used within VoID descriptions to express the origins of data sets. Provenance is considered an important part of Linked Data publication practice and is gaining acceptance. Currently, about 35% of Linked Data sets expose some provenance [61].

In addition to the use of these ontologies, the Linked Data community has concerned itself broadly with three other issues. First is how to associate provenance with groups or sets of triples through mechanisms such as "named graphs" [62]. Indeed, provenance was an original motivation for Named Graphs [63].[12] Second is how provenance should be accessed using existing Web protocols [64]. This includes access to provenance by dereferencing resources [65, 66, 67] and a large amount of work on provenance in conjunction with SPARQL [68, 69, 31, 70, 71]. This issue led to the development of - as a basis for further community harmonization. Third, is the tracking of provenance within the generation of Linked Data, which often is the result of combining or integrating multiple sources [72, 73]. All three of these issues assume the presence of a provenance ontology.

Overall, these ontologies and their use demonstrated the need for a standard for interoperable interchange of provenance. Likewise, they fed into the design process at the start of the overall move towards standardization as discussed in the next section.

4. Provenance Incubator and mapping to OPM

Given the plethora of ontologies for provenance within the Semantic Web community and the community movement that led to OPM, the ground was set for a move towards standardization. At a Dagstuhl Seminar on reflecting on the Semantic Web research after 10 years [74], discussions led to the idea of starting a W3C Incubator Group to investigate potential standardization. At that meeting, Yolanda Gil agreed to chair the group and later with Ivan Herman wrote a charter proposal that was submitted to the W3C. The Provenance Incubator Group was approved in September 2009 and ended in November 2010.

The group performed a use cases and requirements analysis and created a state-of-the art survey, which was subsequently published [25]. To help organize its analysis, the group chose to summarize over 30 use cases it collected into three flagship scenarios. Each scenario presented a situation and then identified associated provenance issues. The three scenarios were:

1. news aggregation, which illustrated how content is aggregated and diffused across the Web;
2. diseases outbreak, which illustrated scientific data analysis and how results are propagated into public policy;
3. business contracts, which looked at issues to do with business process and compliance.

The group used these scenarios to help illustrate a series of requirements for provenance on the Web. These requirements were classified according to 3 categories: content, management and use. The content category refers to what should be contained in provenance data. Management refers to how provenance data should be captured and maintained. Lastly, the use category is about how provenance solutions solve specific user problems. These dimensions helped the incubator group when organizing its state of the art survey. We build on these categories to classify requirements for (see Section 5).

The incubator group also published a report that mapped, using SKOS, many of the ontologies and vocabularies discussed above to OPM [75, 76]. The idea was to understand the commonalities between the existing ontologies and identify, if possible, a common vocabulary within the community. Some of the key findings from the mapping activity that influenced were:

- Many of the ontologies shared the same core concepts, which roughly corresponded to the notions of entities, activities, and agents as defined in OPM.
- There were two main views of provenance, one that was resource-centric and the other more process-centric within in the models.
- Many vocabularies had "shortcut" relationships for modeling common activities. For example, the act of importing a dataset could be modeled as the relation `:data :importedFrom :source`. However, a more extensive description of importing could involve modeling the import activity itself, its length of time, and its inputs (e.g. `:source`) and outputs (e.g. `:data`). Thus, there needs to be a bridge between these two types of modeling approaches.

These items helped shape the construction of the Provenance Working Group charter, which we discuss in Section 6. Beforehand, we propose a categorization of requirements.

5. Categorization of Requirements

To provide some structure to our requirement analysis, we tag each requirement by one or more categories, indicating its

[10] See http://dublincore.org/groups/provenance/

[11] Changeset: http://vocab.org/changeset/schema.html

[12] We note that the Provenance Working Group worked actively with the RDF 1.1. working group to ensure compatibility between and the RDF 1.1. specifications, in particular, with respect to Named Graphs.

broad nature. We refine the Incubator categories [25], *content*, *use*, *management*, but also introduce three further categories *constraints*, *scope*, and *organization*. They are defined as follows.

The *content* category refers to what the data model contains. While standardization avoided restricting specific applications of , some requirements had an impact of how the specifications would be used in practice (these are captured by the *use* category). The *management* category refers to how provenance data should be accessed and packaged up. The *constraints* category refers to requirements that help define semantic grounding and integrity of content. The *scope* category is for requirements that are concerned with the scope of the standardization activity. Finally, the *organization* category encompasses requirements that help give some structure to the specifications.

5.1. Themes and Presentation

Furthermore, requirements have been grouped by section according to the "themes" they related to. Section 6 lists requirements from the incubator group (XG1–XG18). Section 7.1 includes general principles (GE1–GE3). Section 7.2 is concerned with resources (RE1–RE8). Section 7.3 describes the commonly recognized three views on provenance (VI1–VI8). Section 7.4 discusses requirements aimed at making the model usable in practice (EZ1–EZ7). Section 7.5 focuses on the event model underpinning (EV1–EV4). Section 7.6 lists key requirements related to - and - (CO1–CO9). Section 7.7 discusses requirements around provenance of provenance (PP1–PP6). Section 7.8 lists requirements about ontology design (OD1–OD6). Finally, Section AQ1 is concerned with access and querying of provenance (AQ1–AQ4). All requirements, themes and categories are summarized in Table 1. Furthermore, illustrations of the requirements are provided in the form of RDF snippets. A complete description can be found in submitted file `example-expanded.ttl`.

In this article, we distinguish between "initial" requirements (as specified by the Provenance Incubator final report) and "retrospective requirements" (defined in a post-hoc analysis by the authors of this article, based on decisions made along the way and underlying principles emerging from the decisions and design). They will be expressed using the following notation.

Requirement XG*n*. *is to comply with an "initial" requirement, explicity identified by the Incubator Group prior to standardization.*

Requirement GE/RE/VI/EZ/EV/CO/PP/OD/AQ. *is to comply with a "retrospective" requirement, guideline, and design decision, which is formulated in this article and which emerged during the course of the W3C Provenance Working Group.*

Wherever possible, we try to present evidence of these requirements, by referring to Provenance Working Group Resolutions, email discussions, or Wiki pages. They are respectively noted R-⟨*year*⟩-⟨*month*⟩-⟨*day*⟩/⟨*number*⟩, Mail-⟨*topic*⟩, and W-⟨*topic*⟩. These references contain links that are directly clickable in the electronic version of the document.

5.2. Requirements Summary

In this section, we summarize the requirements enumerated in Table 1.

Under the theme "Initial Requirements" (Section 6), we find a focus on interchanging provenance, and a need for multiple serializations of a common conceptual data model, according to users' preferences. Furthermore, several requirements identify core concepts for a standard model of provenance. These include three core notions, resource, activity, and agent, and common inter-relations found in extant provenance vocabularies [75, 76]. Finally, some mechanisms to package up provenance statements, share them, and attribute them are identified as necessary.

Before delving into technical requirements, the theme "General Principles" (Section 7.1) lists broad principles adopted by the working group, such as a commitment to promote usage of the data model rather than restrict its use and to encourage symmetry in the model to facilitate its understanding.

The theme "Resources, Entities and Attributes" (Section 7.2) tackles requirements for the concept of resources, whether mutable or not, and how they should be modeled from a provenance perspective. For this reason, the notion of entity with a fixed set of attributes is introduced. Further requirements are also concerned with a common kind of entity, a collection, which consists of other entities.

The theme "Three Views" (Section 7.3) encompasses requirements related to the three core notions of Entity, Activity, and Agent. They are respectively related to three commonly-encountered perspectives on provenance, namely data flow, process flow, and responsibility in .

A great deal effort has been put to make easy to use, with requirements captured in the theme "Ease of Use" (Section 7.4). They cover: being able to make simple provenance statements; a core for to make it accessible, differentiated from extended parts to cover more complex cases; the choice of namespace; and notational and graphical representations.

In the theme "Event" (Section 7.5), it is explained that is a vocabulary to describe how a system evolved in the past. Requirements are introduced to characterise a system's evolution in terms of events, marking the occurrence of changes pertaining to provenance. Associated with this, is a notion of event ordering, akin to flow of time, but not requiring to make assumptions about clocks.

The promoting of ease of use over the restricting of the vocabulary resulted in a permissive vocabulary. Under the theme "Constraints" (Section 7.6), a set of requirements are concerned with the notion of valid provenance (to be understood as logically-consistent provenance). The ultimate aim is to allow provenance validators to be implemented.

Under the theme "Provenance of Provenance" (Section 7.7), requirements scope a solution to allow provenance of a set of provenance statements to be expressed. In particular, the positioning of with respect to the then-emerging RDF Recommendation (including named graphs) is explored.

Many requirements apply to the conceptual data model in general. However, the theme "Ontology Design" (Sec-

Theme/Section	No	ID	Requirement Title (illustration)	Categories: content	constraints	use	scope	organization	management
6 Initial Requirements	1.	XG1	Interchange				✓		
	2.	XG2	Conceptual Model with Serializations			✓			
	3.	XG3	Resource	✓					
	4.	XG4	Activity	✓					
	5.	XG5	Agent	✓					
	6.	XG6	Generation	✓					
	7.	XG7	Use	✓					
	8.	XG8	Derivation	✓					
	9.	XG9	Version	✓					
	10.	XG10	Ordering of Activities	✓					
	11.	XG11	Association	✓					
	12.	XG12	Time	✓					
	13.	XG13	Location	✓					
	14.	XG14	Role	✓					
	15.	XG15	Plan	✓					
	16.	XG16	Collection	✓					
	17.	XG17	Container	✓					✓
	18.	XG18	View/Account	✓					✓
7.1 General Principles	19.	GE1	Class Disjointness		✓				
	20.	GE2	Mirror		✓				
	21.	GE3	Past (1)	✓					
7.2 Resources, Entities, Attributes	22.	RE1	Entity (2)	✓					
	23.	RE2	Attributes (2)	✓					
	24.	RE3	Non-Characterizing Attributes (2)	✓	✓				
	25.	RE4	Identity (2)	✓					
	26.	RE5	Specialization (3)	✓					
	27.	RE6	Alternate (3)	✓					
	28.	RE7	Collection vs Dictionary	✓					
	29.	RE8	Dictionary Operations	✓					
7.3 Three views	30.	VI1	Three Views	✓					
	31.	VI2	Provenance of Agents (4)	✓					
	32.	VI3	Agent as Entity (4)	✓					
	33.	VI4	Agent as Activity	✓					
	34.	VI5	Derivation is not Transitive		✓				
	35.	VI6	Optional Derivation Path (5)	✓					
	36.	VI7	Activity		✓				
	37.	VI8	No SubActivity	✓			✓		
7.4 Ease of Use	38.	EZ1	Scruffy and Proper (6)	✓		✓			
	39.	EZ2	Separate Vocabulary and Constraints					✓	
	40.	EZ3	Core and Extended Structures					✓	
	41.	EZ4	Common Subtypes					✓	
	42.	EZ5	A Single Namespace			✓		✓	
	43.	EZ6	Layout Convention			✓		✓	
	44.	EZ7	Human Readable Notation			✓		✓	
7.5 Events	45.	EV1	Activity Lifetime		✓				
	46.	EV2	Entity Lifetime		✓				
	47.	EV3	Events Ordering		✓				
	48.	EV4	Instantaneous Events		✓				
7.6 Constraints	49.	CO1	Validity		✓				
	50.	CO2	Equivalence		✓				
	51.	CO3	Constraints Not Specified		✓		✓		
	52.	CO4	Decidability of Validation		✓				
	53.	CO5	ProvRDF Mapping Out of Scope		✓		✓		
	54.	CO6	Alternate Properties		✓				
	55.	CO7	Specialization Properties		✓				
	56.	CO8	Events Preordered		✓				
	57.	CO9	Simultaneous Events		✓				
7.7 Provenance of Provenance	58.	PP1	Provenance of Provenance (11)	✓					✓
	59.	PP2	Named Graph	✓					
	60.	PP3	Bundle (11)	✓					
	61.	PP4	Scope and Nesting	✓					
	62.	PP5	Bundle Name	✓			✓		
	63.	PP6	Bundle Linking	✓					
7.8 Ontology Design	64.	OD1	OWL2-RL Profile		✓				
	65.	OD2	Inverse Relation	✓					
	66.	OD3	Directed Qualified Relation Pattern (5)	✓					
	67.	OD4	Influence (12)	✓					
	68.	OD5	OWL Term Organization					✓	
	69.	OD6	Context for Role (13)	✓					
7.9 Access and Query	70.	AQ1	Reuse Standards						✓
	71.	AQ2	Representation Independence						✓
	72.	AQ3	By Reference and By Value						✓
	73.	AQ4	Services and Resources						✓

Table 1: Categorization of Requirements

tion 7.8) accounts for issues related to the design of an ontology for , some of which in turn influenced the conceptual model.

Finally, in the theme "Provenance Access and Query" (Section 7.9), requirements for making provenance accessible on the Web are discussed.

6. Initial Requirements for

Section 4 discusses the Provenance Incubator Group's critical finding pertaining to a core set of provenance terms that are common across the different provenance terminologies [75, 76]. This finding is quite remarkable: indeed, despite the diverse motivations and perspectives that led to these terminologies, the group was able to establish mappings among them and successfully demonstrate that there are several common concepts in provenance.

In its final report, the W3C Provenance Incubator [9] makes a set of recommendations, identifies priorities, and highlights the importance of standardization of a core set of concepts: it argued that failure to tackle effective standardization in a timely manner could impede effective reuse of open data. Standardization around this set of concepts was auspicious because the field was ripe for immediate progress, thanks to a breadth of expertise and experience and major previous efforts that enjoyed significant uptake. To prepare for standardization, the Incubator Group drafted a charter, setting out the mission, scope, and deliverables of a standardization activity. This draft charter, refined and then approved by the W3C membership, led to the formation of the W3C Provenance Working Group, in April 2011. The rest of this section discusses key aspects of the charter.

The overarching approach adopted by the Provenance Working Group is to consider an (extensible) core provenance language that allows any provenance model to be translated into such a *lingua franca* and exchanged between systems. This is captured by the following requirement.

Requirement XG1 (Interchange)**.** *is to be concerned with the exchange of provenance information.*

Consequently, is not intended to dictate how a system should implement provenance internally. Instead, heterogeneous systems can elect to export their provenance into such a core provenance language, and applications that need to make sense of provenance can then import it and reason over it. This naturally brings the pragmatic question, as to which concrete serialization or format one should adopt to express provenance. Given that is aimed at heterogeneous systems, using multiple, sometimes incompatible, technologies, it was decided that a conceptual data model for provenance was desirable, and it should be serializable in various languages[13], such as Turtle and XML, to facilitate integration with heterogenous systems.

Requirement XG2 (Conceptual Model with Serializations)**.** *is to be defined as a conceptual data model that can be mapped onto various serializable Web languages.*

[13] A serialization to JSON was developed outside the Provenance Working Group [77].

Under the purview of this overarching approach, seven deliverables were identified, which we summarize below.

1. The *conceptual model* specification is a natural language description and graphical illustration of the data model concepts. During the standardization activity, this deliverable took the shape of several documents, including Recommendations: - [1], - [12], - [13] and separate Notes: - [18] and - [19].
2. A *vocabulary* expressing the conceptual model in a Semantic Web language, such as OWL, with a view to map the conceptual model to RDF. This led to: - [14].
3. A *formal semantics* which consists of a mathematical definition of to resolve ambiguities that may arise from the conceptual model specification. This led to: - [17].
4. Web-based protocols to *access and query provenance*. This led to: - [20].
5. A native *XML serialization* of . This led to: - [15].
6. A *primer* is an educational document that provides users with an easy to understand description of the model. This led to: - [11].
7. A *Best Practice Cookbook* is intended to make the link with other relevant notions, such as Dublin Core provenance-related concepts[14]. This led to: - [16].

The *conceptual model* and *vocabulary* deliverables were set to become W3C Recommendations, for which there is a burden of proof of implementability and inter-operability, whereas the other documents became W3C Notes, technical documents without such a requirement but still approved by Working Group consensus. In the process of defining the data model, it was felt that some concepts were not ready for Recommendation status, and therefore were included in separate notes: - [18] and - [19].

The W3C Provenance Incubator final report [9] lists a set of concepts expected to be found in a standard for provenance. We summarize them as requirements for . We refer the reader to the W3C Provenance Incubator final report [9] for illustrations of these concepts in extant vocabularies.

First, three core notions were identified: resources, activities, and agents. They are the foundational building blocks of provenance vocabularies, and they can be linked using various dependencies, for which requirements are also found below.

Requirement XG3 (Resource)**.** *is to model resources, whether mutable or immutable.*

[14] The charter also suggested issues such as licensing in Creative Commons and the OpenId identity mechanism for people, but these were not addressed by the Provenance Working Group. The group also compiled examples of use of provenance https://dvcs.w3.org/hg/prov/raw-file/tip/bestpractices/BestPractices.html, and common questions regarding provenance were answered in a FAQ https://www.w3.org/2001/sw/wiki/PROV-FAQ.

Requirement XG4 (Activity)**.** *is to model executions of computation, whether workflow, program, or service, but also activities in the world, outside computer systems.*

Requirement XG5 (Agent)**.** *is to model humans or other things involved in activities.*

The lifecycle of resources, e.g. when they are created and used, how they are transformed and versioned, is crucial to provenance, as expressed by the following requirements. The community consensus was that the terms generation, use, derivation, and version should be adopted for these notions, respectively.

Requirement XG6 (Generation)**.** *is to model the creation of resources.*

Requirement XG7 (Use)**.** *is to model the usage of resources.*

Requirement XG8 (Derivation)**.** *is to model the derivation of resources from other resources.*

Requirement XG9 (Version)**.** *is to model the versioning of resources.*

While resources are fairly well understood, because they correspond to data or documents, executions are more intangible, because they "happen". Thus, an important aspect of their description is how they relate to each other, who is involved in them, and to what extent.

Requirement XG10 (Ordering of Activities)**.** *is to model how activities trigger other activities.*

Requirement XG11 (Association)**.** *is to model agents participating in activities.*

Note that the incubator had an explicit requirement for a notion of agent *controlling* an activity. The Provenance Working Group opted for a looser notion of association, allowing all the following to be seen as association: a spectator attending a theatre performance, an actor playing in the performance, the director of the show, and the funder for this cultural activity.

There are several additional concepts that are pertinent to provenance, such as time, location, role, and program definition. It was recognized that it is not the purpose of a provenance standardization activity to specify them. Instead, a provenance standard should be able to link to or refer to such concepts, defined elsewhere.

Requirement XG12 (Time)**.** *is to offer the means to refer to time information.*

Requirement XG13 (Location)**.** *is to offer the means to refer to location descriptions.*

Requirement XG14 (Role)**.** *is to offer the means to refer to roles.*

Requirement XG15 (Plan)**.** *is to offer the means to refer to existing description of plans, programs, workflows, or scripts.*

Given the need to deal with both individual resources and sets of them (e.g., data sets or artifact catalogs), the ability to model the provenance of collections was perceived as important. However, it was also acknowledged that such a topic, in itself, is very broad and widely studied, but still involves significant research, for instance, in the database community. Thus, a provenance standard is to incorporate a minimalistic notion of collection, with a focus on their *derivations*. This minimal representation permits users to adopt any extant collections model that suits their needs.

Requirement XG16 (Collection)**.** *is to model a lightweight notion of collection.*

Finally, a provenance language has to provide some "housekeeping" constructs, two of which were identified.

Requirement XG17 (Container)**.** *is to offer a mechanism to package up provenance statements, and present them as evidence for something.*

Requirement XG18 (View/Account)**.** *is to offer a mechanism allowing multiple (possibly different and contradictory) provenance descriptions to co-exist.*

Requirement XG18 is particularly significant. It acknowledges that there may not be a single authoritative source of provenance, and the standard should be architected to accommodate an open view of provenance.

7. Retrospective Analysis for

For expediency, the charter of the Provenance Working Group did not include an explicit deliverable on requirements for provenance. It was then felt that the requirements captured in the W3C Incubator group [25], the W3C Incubator final report [9], previous requirements documents [35] and extensive surveys [5, 4, 78, 2] provided sufficient background and understanding of the field to proceed with standardization. The purpose of this section is to redress this shortcoming, by eliciting, post-hoc, the requirements and the design decisions necessary to make a well-formed and useable set of specifications. For simplicity we refer to all requirements, guidelines and design decisions as requirements below.

7.1. General Principles

To allow design decisions to be made, guiding principles were needed. These took the form of rules coming from the nature of standardization, and softer constraints driven by the desire to ensure the standardization outputs would be adopted and found useful, described below. All the principles were necessarily treated with some flexibility, rather than as absolute obligations.

The fact that was developed as part of a standardization exercise meant that certain principles held: *(i)* Recommendations should not exceed the state of the art, i.e. should not

include new or speculative concepts; and, *(ii)* Recommendations should cover key and common provenance-related concepts from existing provenance models. The fact that Recommendations were developed within the W3C's Semantic Web activity meant that another principle guided the group's decisions: *(iii)* Recommendations should apply to provenance as used in distributed, especially Web-based, settings.

The latter principle was not seen as excluding other domains of use, and another, general principle was observed: *(iv)* Recommendations should not pre-empt the uses to which they will be put and should be applicable to as wide a range of applications as possible. More specific principles then followed from this: *(v)* the recommended models should be general from any given application; *(vi)* the recommended models should be extensible to express the kinds of past occurrence identified in the use cases; and, *(vii)* Recommendations should only include strongly justifiable constraints on how can be used. The last of these principles meant that, in the models being developed, there was a wish to ensure concepts were used for description rather than to restrict what else could be described, leading to the following high-level design decision.

Requirement GE1 (Class Disjointness)**.** *is to minimize class disjointness constraints and to use strong rationale when defining such constraints.*

Another design decision drawn from the desire not to pre-empt use of was based on the observation that many provenance concepts have a complementary 'mirror' concept, e.g. creation is mirrored by destruction, initiation by termination, etc. Even if these mirror concepts are not referred to explicitly in known use cases, their usefulness and relevance to provenance can be predicted, and so should be included in .

Requirement GE2 (Mirror)**.** *is to include the mirror of each concept, where relevant.*

A final consideration was that Recommendations had to balance ease of use with the expressivity needed to cover possible applications. A decision was made to divide the model into two parts: core and expanded. The following principle was then applied: *(viii)* the core model should be easy to apply quickly and without knowledge of the bulk of the recommendations.

Provenance is *not* a workflow language or programming language: provenance is intended to describe what happened, whereas a workflow language is a specification of an execution, which may or may not happen.

Requirement GE3 (Past)**.** *is aimed to describe past executions, as opposed to specify potential future executions.*

A consequence of this is that the Provenance Working Group decided to express influence relations with a verbal form in the *past* (see R-2011-09-01/3[15]) to emphasize that aspect of .

[15]Resolution 2011-09-01/3: http://www.w3.org/2011/prov/meeting/2011-09-01#resolution_3

Illustration 1 (GE3)**.** *properties have a past verbal form.*

```
ex:chart   prov:wasGeneratedBy    ex:compile .
ex:chart   prov:wasDerivedFrom    ex:dataset .
```
□

7.2. Resources, Entities and Attributes

One of the core concepts identified by the Provenance Incubator group was that of resources, which may be immutable or mutable (Requirement XG3). In referring to a mutable resource in provenance, it needs to be clear what state of the resource is intended. For example, consider a Web page of which there were two versions, the first including some claim and the second with the claim removed. If the provenance describes the consequences of agents reading and acting on that claim, then it should refer to the first version of the Web page and not the second, else the provenance will be nonsensical or misleading. It was also noted that the state of a resource did not just include its content, but also context, e.g. the location of the Web page.

One possibility considered was for to model only immutable resources, and require each state to be separately identified (as OPM does). However, this approach was found to have a few problems. First, the provenance would still need to refer to the identified resources of which people wish to describe the provenance, e.g. a Web page identified by its URI, and these are mutable. Second, at least in some cases, it can be impractical to decide whether to model a resource as being in a new state or not, as the context of the resource can itself be defined in different ways. Finally, there are ease of use implications (discussed further in Section 7.4), as each new state requires a new identifier, which is heavyweight when a user wishes to assert a simple statement about their Web page's origins, for example.

Therefore, an alternative approach was taken. It was noted that many changes to a resource's content or context would not have any relevance to the provenance information to be expressed. There are only certain *attributes* of the resource that matter, such as the presence of the claim in the Web page example above. As a first step, a requirement emerged for a concept of a resource that is immutable in certain attributes, which was termed an *entity*.

Requirement RE1 (Entity)**.** *is to model resources with fixed attributes, called entities.*

Activities, agents, and most relations have their own attributes which, similarly to those of entities, can be relevant to what else has occurred as documented in the provenance. The encoding of attributes of relations is described in Section 7.8.

Requirement RE2 (Attributes)**.** *is to model the attributes of entities, activities, agents, and most relations.*

The Provenance Working Group discussed the implications of expressing attributes as part of distinguishing entities. For some general entities, e.g. the Web page above, the only fixed attribute may be the identifier of the page, i.e. its URI, not some additional characteristic. It was decided that it should not be mandatory to express any attribute, even if it was a characterizing attribute. Also, resources will have attributes that are mutable but not relevant for distinguishing between entities in the

provenance, e.g. the background color of the Web page may change but we do not want to document the history of these changes. It was decided that would not define which attributes were fixed and which were not.

Requirement RE3 (Non-Characterizing Attributes)**.** *should allow attributes to be expressed of an entity even when they do not characterize that entity (distinguish it from other entities), and it should be possible to specify entities without requiring characterizing attributes to be expressed.*

In the Web architecture, resources are identified by URIs. Therefore, for compatibility, the following requirement applies.

Requirement RE4 (Identity)**.** *is to use URIs to identify instances of its data model.*

Illustration 2 (RE1,RE2,RE3,RE4)**.** *In Figure 2, the dataset has an identity given by its URI (ex:dataset) and has a further fixed attribute: its title. The dataset title is non-characterising since there may be other datasets with the same attribute.*

```
ex:dataset a prov:Entity;
  schema:headline "Employment Data 2014".
```
□

The concept of an entity allows for both mutable and immutable resources to be modelled. The Web page mentioned above, for example, would be an entity identified by its URI. If a resource never changes in a way that has any relevance to the provenance statements about it, e.g. what is derived from it, then the resource and the entity referred to in the provenance can be one and the same. In other cases, a new entity will have to be identified for each change to the attributes of the resource. Continuing the example above, the Web page with the claim and the Web page without the claim will be separately identified entities, with different attributes (one has the claim, the other does not). However, when a query is made for the provenance of the Web page, the URI of the Web page itself will be used, not the identifier of either more specific entity, which exist purely to document the provenance. In general, a resource needs to be connected to the entities which represent the periods in which that resource had particular attributes. Put another way, a link is required between a specialized entity with a set of fixed attributes and a more general entity with only a subset of those attributes fixed.

Requirement RE5 (Specialization)**.** *is to model the relation between an entity with a set of fixed attributes to a more general entity with only a subset of those attributes fixed, described as the former being a* specialization *of the latter.*

When multiple different parties are documenting the same process, there may be multiple entities that are each views on the same resource, fixing particular attributes relevant to the different provenance statements being made. To make sense of these different views, it is required to relate them, to say that they are both alternative perspectives on the same resource.

Requirement RE6 (Alternate)**.** *is to model the relation between entities that present alternative fixed attribute views of the same resource.*

Illustration 3 (RE5, RE6)**.** *The data set (*`ex:dataset`*) may be a revision of a previous version of the data (*`ex:oldDataset`*). Both versions are a specialization of* `ex:data`*, a data set on employment data, irrespective of its version. Furthermore, each version is an alternate of the other. This is captured by the following RDF triples.*

```
ex:dataset prov:wasRevisionOf ex:oldDataset .
ex:dataset prov:specializationOf ex:data .
ex:oldDataset prov:specializationOf ex:data .
ex:dataset prov:alternateOf ex:OldDataset .
```
□

The Provenance Working Group considered carefully whether new properties were truly needed for the specialization and alternate relationships, or whether existing properties such as `rdf:type`, `rdfs:subClassOf` or `owl:sameAs` could be used instead. As the above illustration suggests, the specialization and alternate properties can relate entities (such as `ex:dataset`) that are "instances" and not necessarily "classes". This distinguishes specialization conceptually from both the `rdf:type` relation that relates an instance to a class, and the `rdfs:subClassOf` relation that relates a subclass to a superclass. Moreover, while `owl:sameAs` can relate arbitrary instances, it is stronger than `prov:alternateOf`: for example, `ex:dataset` may have different values for certain attributes than `ex:oldDataset`. Treating alternate entities as the same would inappropriately collapse distinctions among different versions of the same resource.

As discussed in Section 3.2, collections are important resources in the context of scientific workflows [79] and other domains. This led to a Provenance Incubator requirement on specifying a lightweight notion of collection (see Requirement XG16).

Some preliminary work on collections in OPM [80] modelled collections as entities, to which elements (also entities), can be added or removed, resulting in novel entities. Hence, the adding or removing of elements can be modelled by derivations. With such a modeling, the state of a collection can be inferred, if its initial state is known, and all operations it underwent are known. Working drafts[16] exist illustrating the kind of inferences that may be possible. The specific modelling and axiomatisation that was drafted was using a notion of key to index the elements of the collections.

The Provenance Working Group referred to this type of structure by the term '*dictionary*', while it used the term '*collection*' for the abstract notion of collection, without specific reference to its structure (see D-2012-04-26[17], R-2012-04-19/7[18]). It was recognized that the notion of dictionary was useful, but was

[16] Example of inferences over dictionaries: https://dvcs.w3.org/hg/prov/raw-file/fb00155c3f2e/model/working-copy/wd6-collections-constraints.html, https://dvcs.w3.org/hg/prov/raw-file/7b668ffc729b/model/working-copy/wd6/wd6-collections-constraints.html

[17] Discussion Point 2012-04-26: http://www.w3.org/2011/prov/meeting/2012-04-26#Collections

[18] Resolution 2012-04-19/7: http://www.w3.org/2011/prov/meeting/2012-04-19#resolution_7

only one of the many types of collections that exist (others include arrays, sets, multi-sets, etc). Supporting all of them as part of was not desirable.

Requirement RE7 (Collection vs Dictionary)**.** *is to model a lightweight notion of collection, and only one refinement dictionary, where elements are indexed by keys.*

The topic of collection was hotly debated. In particular, the discussion focused on the key question as to whether the whole collection definition inclusive of dictionaries should be included in Recommendations. Some members felt that collections should not be included as they were not core to the model. Others argued that collections are fundamental to so many domains that they need to be included for interoperability.

Overall, in the spirit of Requirement XG16, the lightweight notion of collections was kept in Recommendations, whereas the more involved notion of dictionary was specified in a separate note (see R-2012-06-22/2[19]). The choice of a Note as a maturity level for dictionaries is in line with the group guiding principle (see Section 7.1) that Recommendations should not exceed the state of the art. Freed from the constraints of Recommendation status, the specification on dictionaries flourished into - [19].

As the discussion of Requirement OD4 shows, there was no consensus to make the general collection membership relation an influence (and specifically a derivation). In contrast, operations over dictionaries are seen as derivations.

Requirement RE8 (Dictionary Operations)**.** *is to model primitive operations over dictionaries as derivations.*

Requirement RE8 was satisfied by introducing Inference D3 (membership-insertion-membership), which makes a dictionary derived from all the members inserted into it.

7.3. Three Views

Depending on their contexts, users may adopt very different perspectives about provenance. Librarians often focus on attribution, i.e. the individuals or institutions who bear responsibility for a given artifact (e.g., author, editor, funder, contributor). Software developers, with version control systems, focus on the versioning of documents, and the derivation of files from others [81]; likewise, data journalists [82] care about primary sources, and intermediary data sets they relied upon. Workflow developers and business analysts have an interest in processes and their inter-relations. These three perspectives are respectively referred to as *responsibility view*, *data flow view*, and *process flow view*.

1. The *responsibility view* is about assigning responsibility for a given result or for what happened in a system.
2. The *data flow view* is concerned with the flow and transformation of information inside computer systems or the transformation of things in physical or imaginary worlds.
3. The *process flow view* is a refinement of the responsibility and data flow views that includes the activities that occurred, which entities they used, how they started and ended, as well as their start and end times.

Requirement VI1 (Three Views)**.** *is to support the responsibility view, data flow view, and process flow view.*

The term 'agent' is overloaded in computer science, carrying different meanings in different communities, as illustrated by the different definitions: `foaf:Agent`[20], (intelligent) agent [83], and (user) agent[21]. Given the desire for to be usable in any application context, it was not considered suitable to prescribe a definition of agent. Instead, an agent is defined by the relation that it is involved in: an agent is responsible for an entity (in that case, the entity is said to be attributed to the agent); an agent is responsible for an activity (in that case, the activity is said to be associated with the agent); and, an agent is responsible for another agent (in that case, the latter agent is said to act on behalf of the former agent).

Given that an agent is to carry responsibility for something (entity, activity, and agent), one needs to be able to talk about the provenance of an agent.

Requirement VI2 (Provenance of Agents)**.** *is to be able to express the provenance of agents.*

This can be addressed by allowing agents to be entities, so that we can use the same modeling constructs to express what they derive from, or their ancestor versions. This leads to the following, more specific, requirement.

Requirement VI3 (Agent as Entity)**.** *is to allow agents to be entities.*

Illustration 4 (VI2,VI3)**.** *In Figure 2, before working for the government, Edith was employed by an IT firm.*

```
ex:edith prov:wasDerivedFrom itfirm:edith.
```
□

Surprisingly, a consequence of Requirement GE1 and Requirement GE2 is that there was no obvious rationale to disallow agents from being activities.

Requirement VI4 (Agent as Activity)**.** *is to allow agents to be activities.*

As a result, being an agent is not an intrinsic characteristic of an entity or activity. Instead, it is the very presence of responsibility relations that implies that some entities or activities are also agents.

As far as the data flow view is concerned, the transformation and the flow of entities is what refers to as a *derivation*. While it is recognized that in some cases specific notions of derivation can be regarded as transitive, there are examples

[19] Resolution 2012-06-22/2: http://www.w3.org/2011/prov/meeting/2012-06-22#resolution_2

[20] foaf:agent http://xmlns.com/foaf/spec/#term_Agent

[21] User agent in http: http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html#sec14.43.

in which this property does not obviously hold[22]. Given this, the Provenance Working Group could not reach consensus on a transitive derivation relation (see ISSUE-612[23]); thus, derivation is not defined as a transitive relation.

Requirement VI5 (Derivation is not Transitive)**.** *is not to mandate derivation to be transitive.*

If users need a notion of transitive derivation, it is still possible to define a subrelation of derivation that is transitive. Or, more simply, derivation may be treated as transitive within particular applications and queries (including SPARQL, using property paths).

To allow for provenance-based reproducibility of results [84], and following some completeness results [85], it is useful to be able to link a derivation with the activity it is underpinned by, and with associated generation and usage events. This extra information associated with derivations is seen as a refinement of derivation useful to support use cases that require more detail.

Requirement VI6 (Optional Derivation Path)**.** *is to allow for derivations to be optionally refined by a specification of a derivation path, including a usage, an activity, and a generation.*

Illustration 5 (VI6,OD3)**.** *In Figure 2, the chart was derived from the data set by the activity* `compile`*. Using the Directed Qualified Pattern (see OD3), the derivation is refined to include this activity.*

```
ex:chart prov:qualifiedDerivation
   [ a prov:Derivation ;
     prov:entity ex:dataset ;
     prov:hadActivity ex:compile ].
```
□

Finally, process flow is represented by activities. An activity represents something that "happened", whereas an entity is a thing, whether real or imaginary. This distinction is similar to that between "continuant" and "occurrent" in logic [86]. For this reason (see Requirement GE1), sets of activities and entities are disjoint, as expressed by the following requirement.

Requirement VI7 (Activity Entity Disjoint)**.** *is not to allow an activity to be an entity.*

The charter identified an initial set of concepts, and made it clear that the Provenance Working Group should not delve into the details of plans and workflows (see Requirement XG15). Furthermore, the charter did not list a notion of subactivity either. The Provenance Working Group considered[24] a notion of subactivity, but did not understand the implication of introducing such a relation to the model. In fact, there was little prior art about this in the provenance community. There was also some concern that specifying such a relation would overlap with some workflow specification initiatives. For this reason, it was decided that a normative definition of such a relation would not be included in .

Requirement VI8 (No SubActivity)**.** *It is not a requirement of to specify a notion of subactivity.*

Instead, the Provenance Working Group suggested[25] that a relation such as `dcterms:hasPart` could be used by applications to model subactivities; applications would be responsible for ensuring its use is consistent with the model.

7.4. Ease of Use

The need to support "widespread publication and use of provenance information of Web documents, data, and resources" [87] was manifested in the idea that should be as easy to use as possible for a wide range of audiences and in particular Web and application developers. This need for ease of use manifests itself in both the guiding principles of as well as requirements that emerged during its specification. In terms of the guiding principles mentioned Section 7.1, two stand out: that Recommendations be applicable to a wide range of applications and that they be usable in a Web-based setting. During the course of the Provenance Working Group, the following requirements emerged.

A key discussion point was the relationship between mutable and immutable resources as discussed in Section 7.2, in particular, around whether would be able to describe mutable resources. It[26] was realized that the problem arose from the need to be able to address two kinds of use cases:

1. the need to make simple provenance statements about resources already on the Web, for example, that a particular blog was attributed to a particular person[27]; and
2. the need to track in a precise fashion (i.e. every version and modification) the provenance of a resource, for example, as generated by a scientific workflow or version control system.

Provenance corresponding to the first use case was termed "scruffy" by the Provenance Working Group, whereas provenance corresponding to the second use case was termed "proper." This dichotomy resembles the neat vs. scruffy debate in AI [88]. However, the Provenance Working Group felt that both use cases were important: indeed, many existing provenance systems already support precise capture of provenance, whereas enabling Web pages to be marked up using was a key part of why the working group was chartered. This led to the following requirement.

Requirement EZ1 (Scruffy and Proper)**.** *should be flexible enough to support both proper and scruffy provenance.*

[22] Example of non-transitivity of derivation: http://lists.w3.org/Archives/Public/public-prov-wg/2011Nov/0191.html

[23] ISSUE-612: http://www.w3.org/2011/prov/track/issues/612

[24] http://www.w3.org/2011/prov/wiki/ResponsesToPublicComments#ISSUE-447_.28subactivity.29

[25] FAQ: http://www.w3.org/2001/sw/wiki/PROV-FAQ#How_can_I_define_a_sub_activity.3F

[26] Meeting minutes F2F2: http://www.w3.org/2011/prov/meeting/2012-02-02

[27] See the following thread for an illustration: http://lists.w3.org/Archives/Public/public-prov-wg/2011Oct/0186.html.

Illustration 6 (EZ1)**.** *The following property is simple to assert, relating two resources, a dataset and a chart, and therefore is regarded as "scruffy".*

```
ex:chart prov:wasDerivedFrom ex:dataset .
```

On the other hand, the qualified derivation of Illustration 5 constitutes "proper" provenance, where the scruffy assertion has been refined with extra information. □

Indeed, the idea emerged that there should be a path that allows provenance to be progressively refined to provide more details. The specialization hierarchy discussed in Section 7.2, derivation refinement (Requirement VI6), and the Directed Qualified Relation pattern in Section 7.8 are examples of constructs that support this refinement.

Furthermore, the scruffy approach was a strong driver in the design of . An approach could have been to identify the various states of resources (and express how they derive from each other) but this would have prevented the expression of provenance with respect to existing mutable resources. For example, writing `:page prov:wasAttributedTo :bob` would first require the identification of the state of the `:page`.[28] Indeed, a totally state-centric approach would have prevented the "shortcut" relationships that were seen in the original provenance vocabularies that fed into the work on .

Another consequence of Requirement EZ1 is that the Provenance Working Group began to think of ways to ease the usage of for the different use cases. To simplify adoption in the scruffy case, it was decided that should provide a vocabulary with minimal constraints on the usage of the terms defined. This adopts the approach used in SKOS of applying the principle of minimal ontological commitment [23] in order to capture the basic informal semantics of provenance and ensure that the use of the language does not cause unexpected outcomes for the user. An example of such an outcome would be transitive implication where none was intended. This is separate from checking whether the provenance expressed in is 'proper'. Both from prior work and discussions in the Provenance Working Group, there was agreement about what would constitute a minimum level of 'proper' provenance. (What forms this level is discussed more deeply in Section 7.6.) These were viewed as constraints on the usage of the vocabulary. An important notion was that users of the vocabulary should not need to have knowledge of the constraints in order to apply . This led to the following requirement.

Requirement EZ2 (Separate Vocabulary and Constraints)**.** *is to a vocabulary and a set of constraints separately.*

With respect to this requirement, an analogy that the group found helpful was to think of the constraints as a definition for developers of a validator whereas the vocabulary was useful for users of terms. Just as there are many users of HTML constructs and few developers of HTML validators, the same would most likely hold for . By separating the definition of a vocabulary and constraints, the Provenance Working Group aimed to make the specifications easier to access for these different user communities.

One of the difficult balancing acts in the design of was the trade-off between defining enough concepts to ensure interoperability, and defining every construct to do with provenance[29]. To achieve this balance, two requirements emerged. The first requirement was the division of the specification into core and extended structures. Core structures are the essence of provenance information and were limited to just the three classes `prov:Entity`, `prov:Activity`, and `prov:Agent` and their interrelationships. In contrast, the extended structures enable more specific uses of provenance with respect to the three views of provenance (Requirement VI1).

Requirement EZ3 (Core and Extended Structures)**.** *is to have a minimal central core with additional extensions.*

The second requirement, to support interoperability, was to introduce some commonly used subtypes of the core concepts. For example, revision and quotation are often used with respect to provenance but are both subtypes of the notion of derivation. Given their wide use, it would be odd not to make these available. Thus, includes one level of subtypes corresponding to these common cases. One key point is that these subtypes are defined with wide applicability — they place few (if any) requirements on the nature of their subtypes or instances. For example, `prov:Plan` is broad enough to include both handwritten baking recipes as well as XSLT scripts on the Web to be considered an instance of the type. This means that users can easily apply these concepts in their own domains without worrying about violating .

Requirement EZ4 (Common Subtypes)**.** *is to provide common classes that are easily extensible.*

Supporting multiple serializations of a single conceptual model resulted in the question as to what namespace(s) to use. Should each individual serialization - , - , - have its own namespace with mappings between them or should a single namespace be used? Similarly, should the extensions to such as - and - be in the same namespace as the other documents? It was chosen to adopt a single namespace (see R-2012-03-29/1[30]). Inspiration for this decision came from two sources:

1. The Architecture of the World Wide Web[31] draws the distinction between a resource, in our case the conceptual model, and its many possible representations (the various serializations).

[28] See http://www.w3.org/blog/SW/2011/10/23/5-simple-provenance-statements/ for examples of these types of expressions.

[29] See http://lists.w3.org/Archives/Public/public-prov-wg/2011Nov/0209.html for a discussion of this issue with respect to the subtyping of `prov:Agent` and the response to Last Call feedback: http://www.w3.org/2011/prov/wiki/ResponsesToPublicComments#ISSUE-520_.28Person.2FOrganization.2FSoftwareAgent.29.

[30] Resolution 2012-03-29/1: http://www.w3.org/2011/prov/meeting/2012-03-29#resolution_1

[31] http://www.w3.org/TR/webarch/

2. The need for developer simplicity when using many different technologies. The simplicity of a single namespace can be effective, as seen with the schema.org vocabulary where only one prefix is required to express structured markup. Note that the namespace prefix is part of the default context of RDFa[32], which means that RDFa users can just use `prov:` without referring to its namespace URI.

Requirement EZ5 (A Single Namespace)**.** *will have a single namespace.*

There were several ramifications of this decision. First, there was the need to verify that using a single namespace worked across and within technologies, in particular, between XML and RDF and within XML. Indeed, supporting multiple XML schemas with the same namespace turned out to be difficult (see ISSUE-608[33]). Similarly, organizing the RDF terms according to the W3C document that introduced them required additional consideration (see Section 7.8 for its solution and rationale). Secondly, it required the use of content-negotiation so that one can get the various representations (OWL2, XML Schema, and HTML) of from its single URI. Finally, it meant that there was a need to provide a unified namespace page[34] that made cross-references across the various definitions residing in each of the specifications.

While in some cases more technically demanding, providing a single namespace achieves two ease-of-use goals:

1. It provides a single point to find all definitions of terms.
2. It decreases the need for developers to worry about supporting mappings between different serializations. For example, one can use the same vocabulary identifiers within an application independently of how the corresponding model is serialized.

The last requirement pertaining to ease of use was the need for a common graphical layout. When discussing provenance or illustrating it, people often draw provenance graphs. Indeed, it is noted that one of the successes of OPM was that it defined a graphical notation for its concepts. To ensure that the notation was consistent not only in the various specifications but also in other types of material (e.g. slides), the group developed a layout convention[35]. Note, that this is a convention (i.e. a suggestion), and not a normative specification.

Requirement EZ6 (Layout Convention)**.** *There should be a single layout convention used throughout specifications.*

Figure 2 adopts this layout convention. It uses blue rectangles, yellow ellipses, and orange pentagons for activities, entities, and agents, respectively. Nodes are organised so that edges all point upwards.

As work on - and - began, its editors identified some informal requirements or expectations on "sensible" inference over provenance data, expressed in terms of a high-level notation. The definition of this notation and descriptions of constraints added to the length and complexity of these documents; in response to this, the Provenance Working Group decided to restructure the first deliverable into three Recommendation-track documents: - , - and - (R-2012-02-23/2[36], R-2012-04-19/1[37]). - was introduced as a notation aimed at human consumption, and was used extensively across the recommendations, particularly - .

Requirement EZ7 (Human Readable Notation)**.** *is to be equipped with a human readable notation.*

Illustration 7 (EZ7)**.** *Below, a few expressions taken from our running example illustrate the - notation. Full example can be found in* `example-expanded.provn`*.*

```
entity(ex:chart)
activity(ex:compile)
wasGeneratedBy(ex:chart,ex:compile)
wasDerivedFrom(ex:chart,ex:dataset)
```
□

7.5. Events

In , activities have a duration in order to reflect the fact that things can occur over a period of time. An option could have been to delimit an activity by a start time and an end time. The intuition would have been that start time should precede the end time of an activity, but for such a precedence to be verifiable, one would need to introduce assumptions about the clocks used to express time, their synchronization, their granularity, and also the clock observer. As the Provenance Working Group opted for a model of provenance without clock assumption, a notion of instantaneous event was introduced instead.

According to - [13], is implicitly based on a notion of events. Five of them are identified: start, end, generation, usage, invalidation. These events are of interest because they mark a "change of state" in the world: an activity is started or ended, an entity is generated, used, or invalidated. These events are used to formulate requirements about the lifetime of activities and entities.

Requirement EV1 (Activity Lifetime)**.** *is to model activities that occur over a period of time, from their start till their end.*

Requirement EV1 adheres with Requirement GE2, since start and end are two mirror events. Requirement GE2 also led to the introduction of the invalidation event, as the mirror for entity generation. This led to the following requirement.

Requirement EV2 (Entity Lifetime)**.** *is to model entities as things that have a lifetime delimited by the entity's generation and invalidation.*

[32] See http://www.w3.org/2011/rdfa-context/rdfa-1.1
[33] ISSUE-608: http://www.w3.org/2011/prov/track/issues/608
[34] See http://www.w3.org/ns/prov.
[35] See http://www.w3.org/2011/prov/wiki/Diagrams
[36] Resolution 2012-02-23/2: http://www.w3.org/2011/prov/meeting/2012-02-23#resolution_2
[37] Resolution 2012-04-19/1: http://www.w3.org/2011/prov/meeting/2012-04-19#resolution_1

These types of events matter because they enable or disable the occurrence of further events. For instance, an entity cannot be used before generation, but it can be after its generation until its invalidation.

Events always involve an activity and an entity. Thus, the start and the end of an activity also involve an entity which triggered that event. Likewise, the generation, usage, and invalidation of an entity also refer to an activity involved in that event.

Each type of event enables or disables the occurrence of specific types of events, as specified by the following requirements.

Requirement EV3 (Events Ordering)**.** *is to model start, end, generation, invalidation, and usage as follows:*

1. *events involving a follow the start of a and precede the end of a;*
2. *events involving e follow the generation of e and precede the invalidation of e;*
3. *usage of an entity by an activity occur between generation and invalidation of the used entity, and between start and end of the activity.*

A natural question that arises from the definition of usage is whether a used entity can be used again, or whether it was consumed, making it non-reusable. The introduction of invalidation addresses this question, since a usage of an entity that makes it non-usable can be modelled by a usage and an invalidation.

An issue that was debated at length is the relation between events and activities. In , activities "occur"; they "do stuff"; they act upon and with entities. Activities are involved in the generation and usage of entities: as indicated above, an event always occurs in the context on an activity. For some application, if it is useful to see the creation of entities as having a duration, this indeed can be modelled by an activity with a duration. However, what one cares about, from a provenance viewpoint, is when the entity is *completely* created and available for usage, which then is referred to as generation. A generation event, or generation for short, is expressed in as a relation between an activity and an entity. This cannot be modelled by an activity (see ISSUE-499[38]). To avoid potential confusion between activity and start/end/generation/usage/invalidation, it is necessary to make it explicit that start/end/generation/usage/invalidation are instantaneous.

Requirement EV4 (Instantaneous Events)**.** *is to be based on a notion of instantaneous event: start, end, generation, usage, invalidation.*

7.6. Constraints

As discussed in Section 7.4, to minimize specifications of constraints in - , - , and - , all constraints were grouped in a single document - . In response to internal reviews (ISSUE-333[39]), a notion of *valid* provenance was introduced (R-2012-06-23/7[40]): it corresponds to the intuition of "proper" provenance, which is to be contrasted to "scruffy" provenance (see Requirement EZ1).

Requirement CO1 (Validity)**.** - *is to define a notion of validity for* .

Requirement CO2 (Equivalence)**.** - *is to define when two valid instances contain the same information.*

- specifies a notion of valid provenance, defined operationally via an algorithm. At a high level, the algorithm proceeds by first *normalizing* a instance by adding missing information through an inference process, then *validating* the normalized instance by checking that various expected properties hold. The constraints are specified in terms of - and - notation.

The Provenance Working Group considered translating constraint validation to other technologies such as RDF/OWL2, and some such translation efforts were carried out by group members, but it was decided to view such translation efforts as implementations of the constraints rather than as material to be standardized (R-2012-09-06/4[41]). Doing so might have several benefits, such as allowing domain-specific refinements of validity, but was placed outside the scope of the Provenance Working Group since the need for this capability was not clear.

Requirement CO3 (Constraints Not Specified)**.** *is not to specify constraints in terms of other Web standards.*

Normalization consists of expanding short forms of - statements to long forms, replacing some optional arguments with new identifiers (existential variables), applying inferences to add new relations to the instance, and applying uniqueness constraints to merge duplicate information or flag inconsistent use of identifiers. Constraint checking takes place on a normalized instance, and involves checking that certain expected properties hold, e.g. that there are no cycles involving strict precedence in the structure of events, that identifiers are used with types that do not violate the (few) disjointness assumptions of , and that other pathological situations do not arise.

Normalization and validity are defined in terms of a well-understood algorithm from database theory called *the chase* [89]. Essentially, the idea of the chase is to apply inference rules or constraints to an instance, making latent information explicit, until no more such applications are possible. If the chase algorithm terminates, it results in a unique *normal form*, which can be used as a basis for further validation and to compare the information content of different datasets. In general, the chase may not terminate, but it was

[38] ISSUE-499: http://www.w3.org/2011/prov/track/issues/499

[39] ISSUE-333: http://www.w3.org/2011/prov/track/issues/333

[40] Resolution 2012-06-23/7: http://www.w3.org/2011/prov/meeting/2012-06-23#resolution_7

[41] Resolution 2012-09-06/4: http://www.w3.org/2011/prov/meeting/2012-09-06#resolution_4

shown that the inferences and constraints provided by satisfy a property called *weak acyclicity*, which suffices to ensure termination [90]. This also ensures decidability of validation and equivalence checking, which the Provenance Working Group agreed was a basic requirement for the constraints (R-2012-06-22/12[42]).

Requirement CO4 (Decidability of Validation)**.** - *should ensure decidability of validation.*

Moreover, while - provides a basic set of constraints that the Provenance Working Group was able to agree are always reasonable, specific applications may wish to check stricter constraints or apply additional inference rules. The mechanism provided by - can be generalized to allow refined notions of validity, though - does not provide an extensible mechanism for specifying such refinements.

In the rest of this section, we summarize some of the main design choices in - , including: the treatment of optional parameters and the decomposition of validation into several stages: (*i*) Applying inferences; (*ii*) Applying uniqueness constraints; (*iii*) Checking typing and impossibility constraints. The topic of checking ordering constraints is discussed in Section 7.5.

Optional parameters. The treatment of optional parameters was a particular area of concern. In - , some parameters may be omitted, while others are required, whereas in the RDF representation (-), by default, all properties can be omitted, but some can be inferred. In both cases, there is a natural question: Does an omitted parameter (or property link) behave as an *unknown* value, or does omission signify *absence* of a value? This distinction is well-explored in the context of data models for (relational) databases: the semantics of NULL values has been studied extensively, with both unknown-value and missing-value semantics [91].

- formalizes the behavior of optional parameters in - . Optional parameters can arise in two ways in - : via shortened, convenience forms of relations, or via explicit use of a "null" symbol (the special - token −). The shortened forms are expanded to relations that contain all parameters, by inserting − values for missing parameters. Then, optional parameters that are viewed as denoting unknown values are dealt with via definitional expansion, by introducing fresh names for the unknown values. These names are viewed as existential variables, which can potentially be resolved to other identifiers later through *merging* resulting from uniqueness constraints. Optional parameters that carry missing-value semantics are left as − values; such values are viewed as distinct from ordinary identifiers.

The application of this behavior to other representations was not specified; mappings between - and other representations were not formally specified either, although informal descriptions of these mappings were maintained (and considered important as internal documentation) during the Provenance Working Group activity on the W-ProvRDF[43] wiki page.

Requirement CO5 (ProvRDF Mapping Out of Scope)**.** *x is not to formally specify the mappings between different serializations such as - , - and - .*

Inferences. In - , inferences are rules that specify that additional relations can be added to the instance, whereas constraints are rules that check the consistency of information already in the instance (possibly including information added through inference). This difference in terminology is primarily for expository purposes; there is no logical distinction between inferences and constraints, since one can view constraints as inferences whose conclusions are logical falsehood or other auxiliary formulas.

We will not describe all of the inferences in detail, but mention two groups that involve subtle issues. First, we consider inferences that state that any entity has a generation and invalidation event, and that any activity has a start and end event. At one stage in the development of , these inferences were formulated in a way that could lead to an infinite chain of reasoning: any entity has a generation event, which involves some activity, which has a start event, which involves some entity, and so on (ISSUE-465[44]). This potential nontermination was resolved by weakening these inferences to only apply to entities or activities that are explicitly declared (using `entity()` or `activity()` relations). Moreover, care was taken to avoid inferences that introduce new entity or activity declarations. This is why typing constraints (discussed later in this section) do not generate new `entity()` or `activity()` relations, but instead only check that the identifiers involved can be assigned appropriate types.

The second group of inferences that merits discussion concerns alternate and specialization. The Provenance Working Group reached consensus on these relationships only after extended discussions of their possible meanings (W-SpecializationAlternateDefinitions[45]). The formal semantics (discussed later in this section) played an important role in the discussion that led to the adoption of these definitions and associated inferences and constraints, particularly the role and properties of alternate and specialization:

Requirement CO6 (Alternate Properties)**.** - *is to ensure that alternate is an equivalence relation.*

Requirement CO7 (Specialization Properties)**.** - *is to ensure that the specialization relation is an irreflexive partial order and a subrelation of alternate.*

It is important to reiterate that the alternate relation is mathematically an equivalence relation, but it is not `owl:sameAs`. The `owl:sameAs` relation also happens to be an equivalence

[42] Resolution 2012-06-22/12: http://www.w3.org/2011/prov/meeting/2012-06-22#resolution_12

[43] WIKI ProvRDF: http://www.w3.org/2011/prov/wiki/ProvRDF

[44] ISSUE-465: http://www.w3.org/2011/prov/track/issues/465

[45] WIKI SpecializationAlternateDefinitions: http://www.w3.org/2011/prov/wiki/SpecializationAlternateDefinitions

relation, because it indicates that the resources identified by two identifiers are one and the same (and thus exhibit all properties asserted about each). Therefore, `prov:alternateOf` can be used in situations where `owl:sameAs` is inappropriate, for example to link different entities that present different aspects of a common thing from different perspectives, at different times, or from different data sources. Similarly, the `prov:specializationOf` relation can be used to link more specific alternate entities to more generic ones.

Illustration 8 (CO6,CO7)**.** *Continuing with the revision and specialization relationships in Illustration 3, we have:*

```
ex:dataset prov:wasRevisionOf ex:oldDataset .
ex:dataset prov:specializationOf ex:data .
ex:oldDataset prov:specializationOf ex:data .
```

- *specifies that specialization and revision relationships imply alternate relationships, so the following relationships are inferred by normalization, along with symmetric versions of these facts.*

```
ex:dataset prov:alternateOf ex:oldDataset .
ex:dataset prov:alternateOf ex:data .
ex:oldDataset prov:alternateOf ex:data .
```

□

Constraints and validation. Once a instance has been normalized, it can be validated by checking certain constraints, including ordering of events, typing, and impossibility constraints. Of these, the ordering constraints are representative of the design choices and retrospective requirements for constraints and validation. The ordering constraints collect ordering relationships among events; for example, an entity's generation precedes all other events involving it and an activity's end must follow all other events involving the activity (see Requirement EV3). The inferred ordering relationships can be strict, meaning the two events involved must be distinct, but in most cases event ordering relationships allow the two events to be simultaneous without being equal.

Requirement CO8 (Events Preordered)**.** - *is to allow events to form a preorder (not necessarily a partial order). That is, event ordering is transitive and reflexive, but it is possible for two different events to occur simultaneously.*

Illustration 9 (CO8)**.** *In - , the only strict ordering relationship between two events is derivation. Thus, if we consider our running example, it would become invalid if we added any one of the following relationships:*

```
ex:dataset prov:wasDerivedFrom ex:chart .
ex:publish prov:wasStartedBy ex:chart .
ex:publish prov:used ex:chart .
```

The reason is (intuitively) that these relationships would introduce a directed cycle into the event preorder relation, and such a cycle would involve a derivation step, which is not allowed. In contrast, all of the following relationships could be asserted without damaging validity.

```
ex:publish prov:wasInformedBy ex:compile .
ex:compile prov:wasStartedBy ex:chart .
ex:government prov:actedOnBehalfOf ex:edith .
```

The Provenance Working Group did not reach consensus that cycles involving any other relationship besides derivation should be forbidden. Instead, all of the instantaneous events along such a cycle are regarded to be simultaneous. Of course, particular applications are free to impose stricter notions of validity, for example to rule out an entity starting its own generating activity. □

At one stage, the Provenance Working Group considered a stronger constraint (similar to a constraint in OPM) requiring that an entity have at most one generation or invalidation event, and likewise for activities and start or end events. The Provenance Working Group debated this issue and concluded that it was too strong, since it would rule out describing situations in which a composite activity and a component of the activity both (simultaneously) contributed to the generation of an entity (ISSUE-473[46]). Instead, a weaker constraint was introduced requiring that all generation events for a given activity all occur simultaneously.

Requirement CO9 (Simultaneous Events)**.** - *is to require multiple generation events of the same entity to occur simultaneously; similarly for invalidation, start, or end events.*

This issue was discussed fairly late in the development of - . It illustrates the general rules the group adopted for agreeing on constraints: a constraint or inference must have a plausible motivation, must have no intuitive counterexamples, and must be implementable within a decidable formalism (R-2012-06-22/12[47]). Controversial constraints were either dropped (to avoid prematurely standardizing overly-strong constraints) or weakened to avoid the controversial scenarios.

Illustration 10 (CO9)**.** *Consider again our running example. We might also wish to express that the government published the chart as part of a monthly data release. In this case, the chart has two generation events, which we might want to name as* `gen1` *and* `gen2`*, here expressed in - :*

```
wasGeneratedBy(gen1;ex:chart,ex:compile)
wasGeneratedBy(gen2;ex:chart,ex:februaryDataRelease)
wasAssociatedWith(ex:februaryDataRelease,ex:government)
```

This is allowed, but the two generation events are considered to be simultaneous; if this is not intended, then separate entities are needed to disambiguate the chart compiled by Edith and the one incorporated into the February data release. □

Semantics. Developing a formal semantics was an optional goal of the Provenance Working Group charter, and its scope was left unspecified. A draft semantics was maintained on the W-FormalSemantics[48] wiki page and discussed at a Dagstuhl

[46] ISSUE-473: http://www.w3.org/2011/prov/track/issues/473

[47] Resolution 2012-06-22/12: http://www.w3.org/2011/prov/meeting/2012-06-22#resolution_12

[48] WIKI FormalSemantics: http://www.w3.org/2011/prov/wiki/FormalSemantics

seminar in February 2012 [92] (roughly halfway through the Provenance Working Group's lifetime). The goal of the semantics was to capture some of the informal discussion concerning entities, activities, and events, in order to elucidate controversial relationships such as specialization and alternate and their properties. This discussion informed subsequent development of the constraints and informal understanding represented in the other recommendations, leading to consensus on the behavior of alternate and specialization (R-2012-05-03/2[49]).

As noted above, - draws upon background in logic and database theory, such as the chase and weak acyclicity [89, 90]. However, in order to keep it accessible to developers, the WG decided to present the constraints in a way that was intended to appeal to potential validator developers, emphasizing operational aspects (how to check the constraints) over formal or logical aspects (what the constraints really mean). Moreover, - was intended to be self-contained as a specification, and therefore did not rely upon (or heavily cross-reference) external sources for concepts in logic; this also led to the possibility for confusion where the Provenance Working Group adopted notation or terminology different from conventional logical terms. For example, the term "validity" used in - is closer to what logicians would call "consistency", if one views a instance as a logical theory; we chose to use the term "validity" due to its analogous use in other W3C standards. Some public feedback on the constraints amplified the need to explain the relationships and differences between the terminology used in - and that used in logic. In particular, public feedback (ISSUE-576[50]) highlighted the potential problem that - might overspecify constraint checking, by describing an algorithm rather than defining what it means to be valid (ISSUE-581[51]).

While the Provenance Working Group felt that it was preferable for - to present an operational approach in order to increase accessibility to developers, it also agreed with the goal of providing a declarative specification that can be implemented in many different ways. Thus, - explicitly specifies that any implementation that provides the same results as the validity-checking algorithm is compliant. However, the constraints did not provide a high-level, declarative description of validity separate from the algorithm. Instead, the Provenance Working Group ultimately decided to publish this declarative specification as part of a revised version of the formal semantics, - .

In particular, - reviews standard concepts and terminology from logic, explains how they are related to the notation used in - , and gives a corresponding mathematical model. For example, all of the constraints and inferences are restated in - as first-order formulas. In addition, a mathematical model is presented and each relation is assigned a meaning in the model. Every such formula is shown to be sound for reasoning about the proposed class of models; moreover, it is shown that any valid instance has a model (a weak form of completeness).

7.7. Provenance of Provenance

As far as the state of the art was concerned, notions of view over provenance [37] and a notion of account [8] were addressing, in part, the Incubator's requirement XG18 on Views and Accounts. At the same time, the RDF Working Group was actively debating the notion of named graph (see M-2011Feb/0092[52]), distinguishing containers (g-box), from snapshots (g-snap), from their serializations (g-text). It was unclear whether OPM accounts were meant as a container mechanism or a snapshot, and the Provenance Working Group was on the verge of researching the topic, rather than standardizing best practice.

Hence, following multiple discussions (see W-Accounts[53] and W-Graphs[54]), the Provenance Working Group identified the primary requirement for this functionality (see D-2012-02-02[55]) as being able to express the provenance of provenance.

Requirement PP1 (Provenance of Provenance)**.** *is to offer a mechanism to express the provenance of provenance.*

Furthermore, implictly, the Provenance Working Group sought to remain compatible with RDF Named graphs as they were being designed.

Requirement PP2 (Named Graph)**.** *Provenance of provenance is to be expressible using RDF named graphs.*

Since RDF 1.1 was still under development, and therefore not normative yet, the Provenance Working Group did not provide any example of provenance of provenance using named graphs.

Based on Requirements XG18, PP1, and PP2, the Provenance Working Group decided on a *bundle* construct that allows a set of provenance statements to be named. Having a name, one can describe it as an entity, and express its provenance by reusing the existing constructs.

Requirement PP3 (Bundle)**.** *is to model a notion of bundle as a named set of provenance statements.*

Illustration 11 (PP1, PP3)**.** *Our running example, assumed to be denoted by* `ex:example-expanded`*, is a bundle of statements that can be attributed to the authors of this article.*

```
ex:example_expanded a prov:Bundle, prov:Entity ;
   prov:wasAttributedTo ex:Luc, ex:Paul,
                        ex:James, ex:Tim, ex:Simon .
```
□

Following Requirements RE4 and PP2, bundles do not provide a scoping mechanism for identifiers; further, bundles are not to be nested.

[49] Resolution 2012-05-03/2: http://www.w3.org/2011/prov/meeting/2012-05-03#resolution_2
[50] ISSUE-576: http://www.w3.org/2011/prov/track/issues/576
[51] ISSUE-581: http://www.w3.org/2011/prov/track/issues/581
[52] Mail 2011Feb/0092: http://lists.w3.org/Archives/Public/public-rdf-wg/2011Feb/0092.html
[53] WIKI Accounts: http://www.w3.org/2011/prov/wiki/Using_named_graphs_to_model_Accounts
[54] WIKI Graphs: http://www.w3.org/2011/prov/wiki/Using_graphs_to_model_Accounts
[55] Discussion Point 2012-02-02: http://www.w3.org/2011/prov/meeting/2012-02-02#PROV__2d_DM

Requirement PP4 (Scope and Nesting)**.** *is not to allow nesting of bundles and scoping of identifiers.*

In the spirit of compatibility with RDF Datasets[56], the Provenance Working Group did not specify what resource a bundle name is expected to denote.

Requirement PP5 (Bundle Name)**.** *is not to specify what a bundle name denotes.*

However, a linked data approach as adopted by Moreau and Groth [22] suggests that dereferencing a bundle identifier results in a bundle.

As the Provenance Working Group was specifying the bundle construct and as deployment of bundles on the Web was being envisaged, it became clear that bundles would constitute islands of provenance information that would be distributed across the Web. Furthermore, as creators of provenance slice their provenance in bundles, so as to be able to assert their provenance, a further requirement emerged of being able to identify a bundle in which further provenance information can be found about an entity or activity. In applications where provenance is created by multiple parties over time, it is useful for provenance descriptions created by one party to link to provenance descriptions created by another party. Such a mechanism would allow the "stitching" of provenance descriptions together.

Requirement PP6 (Bundle Linking)**.** *is to provide a mechanism for linking entity descriptions across provenance bundles.*

To address this requirement, the group considered a notion of *provenance locator* [57], a data model construct that indicates where, and in which bundle, an entity's provenance can be found (this construct was inspired by `prov:has_provenance`, see Section 7.9). The group was not supportive of making the mechanism for accessing provenance explicit in the data model. Instead, relations such as `sioc:topic`, `foaf:primaryTopic` were considered to express that some bundle contained descriptions about an entity, meaning that this entity was a topic in that bundle. As these relations seem to address part of the requirement, the focus then moved on to the more granular relation that was required between two entities in separate bundles (one "local" and one "remote"). It was felt that it was not appropriate for the Provenance Working Group to introduce a further relation between entities, given the existence of `prov:specializationOf` and `prov:alternateOf` (see D-2012-05-31[58]). As a result, the group opted for a subrelation of specialization, and defined the notion of *mention* that is treated in its own Note [18]. It was recognized that the concept Mention was experimental, and for this reason was not defined in recommendation-track documents (see R-2012-11-09/4[59]).

7.8. Ontology Design

OWL2 Profile. While encoding the conceptual model in OWL2, the Provenance Working Group chose (see R-2011-07-07/6[60]) to design a lightweight vocabulary, with a view to support the linked data approach [72]. This issue was debated at length (see D-2012-02-02[61], M-OWL2-RL[62]), and led to a further decision to settle on the OWL2-RL profile [93], since it is aimed at applications that require scalable reasoning without sacrificing too much expressive power.

Requirement OD1 (OWL2-RL Profile)**.** *The ontology is to be compatible with the OWL2-RL profile.*

Only five axioms of the ontology do not suit the OWL2-RL profile (see [14][63]). All these axioms use an anonymous class union for the domain or range of a property, while OWL2-RL requires the classes to be named explicitly. Their presence is simply ignored by OWL2-RL reasoners, and would thus allow a more permissive domain or range for the property. Although introducing named "placeholder" classes would have suited the OWL2-RL profile, these additional classes would have been a distraction from the core model elements. The non-compliant axioms were thus accepted in favor of ease of use and interoperability with the conceptual model.

Inverses. The core of - (see Section 7.2) is intentionally kept simple to ease the creation of RDF triples, and therefore to promote adoption and maximize interoperability. For one, - avoids introducing too many properties' inverses. While it is logically equivalent to assert either `:e1 prov:wasDerivedFrom :e2` or its inverse `:e2 prov:hadDerivation :e1`, practically, developers consuming both forms of assertion may need to exert extra effort such as adding an OWL reasoner or doubling the size of code and queries to handle both cases. To avoid this extra effort, - promotes[64] most properties over their inverse, so that authors and consumers may focus on one.

Requirement OD2 (Inverse Relation)**.** *The ontology is to define all, but encourage use of certain, property inverses.*

By convention, the preferred property is the one that points "into the past". It is important to note that all property inverses are fully defined, but omitted from the OWL encoding (see Abstract[65] [14]). All preferred properties are annotated with the

[56] RDF Datasets: http://www.w3.org/TR/rdf11-concepts/#section-dataset

[57] Draft: https://dvcs.w3.org/hg/prov/raw-file/7b668ffc729b/model/working-copy/wd6/wd6-bundle.html

[58] Discussion Point 2012-05-31: http://www.w3.org/2011/prov/meeting/2012-05-31#Provenance_Locator___28_hasProvenanceIn__29_

[59] Resolution 2012-11-09/4: http://www.w3.org/2011/prov/meeting/2012-11-09#resolution_4

[60] Resolution 2011-07-07/6: http://www.w3.org/2011/prov/meeting/2011-07-07#resolution_6

[61] Discussion Point 2012-02-02: http://www.w3.org/2011/prov/meeting/2012-02-02#Comments_from_Ivan

[62] Mail OWL2-RL: http://lists.w3.org/Archives/Public/public-prov-wg/2012Feb/0478.html

[63] PROV-O OWL2-RL: http://www.w3.org/TR/prov-o/#owl-profile

[64] PROV-O Inverse: http://www.w3.org/TR/prov-o/#inverse-names

[65] PROV-O OWL file: http://www.w3.org/ns/prov-o

local name of their inverse, should a developer wish to use the inverse instead. The inverses are also enumerated in the Recommendation and defined in a separate OWL document (see Appendix B[66] [14]).

Qualified Relation Pattern. Despite the desire for simplicity, binary relations are not always sufficient to describe situations: for example, a user may want to indicate the time at which an entity was generated by an activity, or they may want to specify the activity for which a delegation of agent responsibility took place. Because these n-ary forms were part of the model, it was essential that - support both. The *Qualified Relation pattern* [94] is a common mechanism to reify binary relations, and provided a basis for design. Because binary relations in have a preferred direction (Requirement OD2), and the Qualified Pattern does not naturally indicate direction, it was important for the Provenance Working Group to evolve the Qualified Pattern into the *Directed Qualified Relation Pattern*. In the former, the qualification instance "points" to each component of the relation that is being described. For example, a qualification for "Marriage" will point to each spouse involved in addition to providing details about the spouses' relationship. In the latter, the subject of the unqualified relation points to the qualification, and the qualification in turn points to the unqualified relation's object while also providing additional details about the relation[67].

Requirement OD3 (Directed Qualified Relation Pattern)**.** *The ontology is to adopt the directed qualified relation pattern to express n-ary relations.*

Within this pattern, binary relations are referred to as *unqualified relations*, and the application of the pattern onto an unqualified relation results in a complementary *qualified relation*, which are viewed as "paralleling" the unqualified relation. The RDF triples of a qualified relation intentionally "flow" in the same direction as the unqualified RDF triple.

The Directed Qualification Pattern has an unstated correspondence to Reification [95]. The `prov:Influence` class is a subclass of `rdf:Statement`; the "prov:qualifiedX" properties are inverses of `rdf:subject`; the subtype of `prov:Influence` implies the value of `rdf:predicate`; and the properties `prov:entity`, `prov:agent`, and `prov:activity` are subproperties of `rdf:object` with ranges specific to - .

As the Directed Qualified Relation Pattern was being deployed across the ontology, it became clear that introducing some structure to the ontology would be beneficial. Hence, a novel qualification, named *Influence* was introduced as a device to abstract from the various Qualifications `prov:Generation`, `prov:Invalidation`, `prov:Communication`, `prov:Delegation`, `prov:Association`, `prov:Attribution` `prov:End`, `prov:Start`, `prov:Usage`, `prov:Derivation`. It carries the idea that there is some form of influence between two resources (R-2012-06-22/6[68]). This relation was not expected to be asserted in descriptions because it is broad. Instead, one of the ten Qualifications should be used; in that sense, the influence relation is "abstract". However, this relation was believed to be useful to express queries. Further, it was deemed useful not only for the ontology, but also for the model as a whole. Thus, the following requirement for .

Requirement OD4 (Influence)**.** *is to model an "abstract" notion of influence.*

Illustration 12 (OD4)**.** *The following SPARQL query shows all influences that led to the chart; it assumes that RDFS reasoning has been enabled.*

```
select  ?y
where ex:chart prov:wasInfluencedBy ?y
```
□

There was no consensus in the Provenance Working Group to consider the following relations as a form of influence: `prov:hadMember` (see R-2012-07-12/1[69]) `prov:specializationOf`, `prov:alternateOf`. Hence, they remained exclusively binary and unqualifiable.

Organization. Grouping OWL terms became necessary as other documents neared completion. The - , - , - , and - notes all introduced new terms that required an OWL representation, but were not Recommendations and thus not part of - . Because W3C Recommendations are fundamentally different from Notes with respect to what must be implemented, it was important to provide these terms in groups that could be accessed separately.

Requirement OD5 (OWL Term Organization)**.** *All terms, from both Recommendations and Notes, are to be defined in OWL.*

Namespaces could not be used to group terms because of Requirement EZ5, which also implied that all terms would be accessible from the single namespace. The solution[70] was to create six ontologies within the base `http://www.w3.org/ns/` that would be combined into a seventh composite ontology `prov#`; the six ontologies were `prov-o#`, `prov-o-inverses#`, `prov-aq#`, `prov-dictionary#`, `prov-links#`, and `prov-dc#`. Although all terms share the same namespace, they appear in different component ontologies and each term uses the `rdfs:isDefinedBy` property to indicate the component ontology that it is in. Finally, the `prov#` ontology `owl:imports` each component ontology, and the component ontologies are included directly so that clients do not need to perform the imports themselves. The `prov#` ontology also reports that it was

[66] PROV-O Inverses: `http://www.w3.org/ns/prov-o-inverses`

[67] Directed Qualification Pattern is illustrated at `http://www.w3.org/TR/prov-o/#qualified-terms-figure`

[68] Resolution 2012-06-22/6: `http://www.w3.org/2011/prov/meeting/2012-06-22#resolution_6`

[69] Resolution 2012-07-12/1: `http://www.w3.org/2011/prov/meeting/2012-07-12#resolution_1`

[70] The OWL ontology design is documented in the FAQ at `https://www.w3.org/2001/sw/wiki/PROV-FAQ#The_PROV_URIs`

derived from (in the sense of prov:wasDerivedFrom) each of the component ontologies, since it already includes them in its representation.

Roles and Locations. The individuals listed on the front page of this article are its authors, whereas the same individuals edited some specifications, or contributed to others. Likewise, a PNG file may be input to a conversion library to JPG, whereas "55" may be a compression rate parameter to this functionality. Author, editor, contributor, input file, parameter are *roles* that some agent or entity can assume in some context (see Requirement XG14).

It should be noted that the concept of role is extensively debated in knowledge representation and ontology design communities. Therefore, since the Provenance Working Group did not want to impose any structure or any prescriptive semantic meaning on roles, anything can be regarded as Role from a perspective.

However, the question that needed to be addressed is what the placeholders for roles are in the data model. Specifically, if roles appear to be meaningful for some context, what should these contexts be? Two contexts were considered by the group.

The context of a role could have been a relation. For instance, an article was attributed to an agent, who acted in some role, e.g. author. Given that roles may apply to agents or entities, roles therefore could apply to either the subject or the object of an attribution relation (or both). This made the expression of roles burdensome, ambiguous, and not natural, and the group failed to reach consensus on an elegant definition R-2012-06-07/2[71].

Alternatively, the context of a role could be an activity. Hence, "55" is an entity that is a parameter in a context that involves that entity and an activity: for instance, in the conversion to JPG. This option was preferred for its simplicity, and led to the following requirement.

Requirement OD6 (Context for Role)**.** *is to define a role, as the function of an entity or agent, in the context of an activity.*

Hence, roles apply to agents and entities in the context of relations involving an activity: namely, these are usage, generation, invalidation, association, start, and end, but no other relation.

Illustration 13 (OD6)**.** *In the following RDF snippet, the role of* `ex:dataset` *is specified to be* `ex:inputDataRole`*.*

```
ex:compile prov:qualifiedUsage
 [ a prov:Usage ;
   prov:entity ex:dataset ;
   prov:role ex:inputDataRole ; ] .
```
□

Likewise, location (see Requirement XG13) is a valuable piece of information, part of the provenance of some resource. As for role, is agnostic about how locations are expressed. Instead, the Provenance Working Group focused on defining the placeholders for location. It was agreed that anything that can be explicitly or implicitly linked with time, can also be provided with a location attribute. This includes entity, activity, and agent, but also relations such as usage, generation, invalidation, start, and end.

7.9. Provenance Access and Query

The aim of - [20] was to provide support for the discovery and accessing of provenance. One of the key issues that arose early in the design process was the concern that the Provenance Working Group would "reinvent the wheel" by specifying a provenance specific access mechanism where already existing Web standards (e.g. SPARQL or resource lookup) could be used. To prevent this, the following requirement emerged.

Requirement AQ1 (Reuse Standards)**.** - *should reuse existing standards and follow Web Architecture principles.*

Meeting this requirement was helped by the discussion in the Provenance Incubator Group about provenance in the World Wide Web architecture[72]. The resulting specification combined existing Web Standards to facilitate access to provenance only adding a few items (e.g., specific link headers) where necessary.

An often discussed concern was what representation - would recommend for provenance data accessed by the protocol (see ISSUE-428[73]). Would the protocol require Turtle, XML, etc. This was a trade-off between encouraging interoperability and spreading adoption. Since it was not guaranteed that any single representation for serializing would be widely adopted, it was decided that the protocol should remain representation agnostic.

Requirement AQ2 (Representation Independence)**.** - *should be independent of a representation.*

Here, another piece of Web architecture, namely, content-negotiation was relied upon in order to deal with the multiplicity of representations.

Within the Provenance Incubator Group, when discussing accessing provenance, a key distinction arose, whether to embed provenance within a document or instead store it externally (e.g., in a provenance store or in a file). This distinction became known as accessing provenance by Reference or by Value. Use cases for both access approaches were given. For example, it might be useful to embed small amounts of provenance within an image file for easy exchange, while if large amounts of provenance are associated with many documents, it is useful to use a dedicated provenance storage facility. Given these use cases, the Provenance Working Group decided to support both access approaches.

Requirement AQ3 (By Reference and By Value)**.** - *should support the access of provenance, both by linking to it (i.e. by reference) and by inclusion within a resource (i.e. by value).*

[71] Resolution 2012-06-07/2: http://www.w3.org/2011/prov/meeting/2012-06-07#resolution_2

[72] Provenance and Web architecture: http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/#Provenance_in_Web_Architecture

[73] ISSUE-428: http://www.w3.org/2011/prov/track/issues/428

The by value case is supported, simply, through standard metadata embedding, for instance, by using RDFa.

To support the by reference case, - specifies a new link header and associated property definition, `prov:has_provenance`, that allows one to point to the provenance information for a particular resource stored at an external location. Associated with `prov:has_provenance` is the definition of an anchor parameter which allows one to find the entity within the provenance corresponding to the resource. A particular point of discussion was around the meaning of multiple `prov:has_provenance` anchor pairs (see M-2013 Feb/0051[74]). When using HTTP headers the pairing is one-to-one, each anchor corresponds to one `prov:has_provenance` link. However, when using the the link definition within HTML there is not a one-to-one pairing. Thus, in the case of multiple `prov:has_provenance` links, the application is required to look through all the provenance information referred to in order to find the anchor resource. The decision to adopt this approach was made in light of Requirement AQ1 to reuse existing capabilities, in this case, the already existing HTML link element.

Related to the notion of provenance being stored by value or by reference, was whether provenance would be hosted as a Web Service or as a Web Resource (see for instance ISSUE-425[75]). Again this led to the requirement to support both styles of interaction.

Requirement AQ4 (Services and Resources)**.** - *should provide support for accessing provenance hosted as a Web resource and through Web Services.*

For the case of the Web Service, it was decided to not overspecify the service definition but allow for extensibility.

One lesson learned from these unwritten requirements is that, in the absence of much prior work, leveraging existing standards and focusing on adoption can lead to a simplified and usable specification.

8. Outstanding Issues

was specified by the Provenance Working Group over the course of two years of activity, with a specific charter that set the scope of its work. Ideas that emerged but were not prime for standardization were included in notes, or simply not pursued at all. This section summarizes some issues that future standardization activities may focus on.

8.1. Model Refinement

The model with its three views offers a good compromise reflecting current practice in pre-existing solutions. While still preserving the requirement of core and extended structure (Requirement EZ3), the mirror principle (Requirement GE2) could be applied more aggressively.

For instance, allows derivations to be refined by making the derivation path explicit, involving a generation, an activity, and a usage (see Requirement VI6). The same pattern does not hold for communication: in a mirror design, communication could also be refined by making the communication path explicit, with a usage, an entity, and a generation. Likewise an attribution could be refined by an attribution path involving a generation, an activity, and an association.

While the notion of fixed attribute is critical in the definition of entities (see Requirement RE1), offers no mechanism to assert which attributes are supposed to have a constant value during the lifetime on an entity, or those that may change. If true discoverability and processing of unknown provenance is to be supported, this information needs to be expressed explicitly.

As noted in Requirement VI8, does not standardize on a subactivity relationship, but it is suggested that similar terms from other vocabularies can be used. Future versions of could standardize this relationship if there is a clear need.

8.2. Validation

- provides a basic set of constraints that the Provenance Working Group was able to agree on as reasonable, and - gave a lightweight formal justification in the form of soundness and weak completeness results. This principled approach should help provenance designers to express valid provenance, and validator implementors to conceive efficient and scalable solutions. However, further formal justification for validation (such as a stronger form of completeness or more intuitive semantic properties) would be desirable for guiding development of vocabularies or future versions of . For example, completeness [85], causality [96], and reproducibility [84] have been studied for previous models such as OPM, and these techniques could be extended to . In addition, the constraints were designed with maximum general applicability in mind, but experience gained in specific contexts such as scientific workflows, business processes, and database queries may motivate additional research on validation.

8.3. Security Aspects

While some specifications briefly discuss security aspects (see - [12] Section 6 Media Type, and - [20] Section 6 Security Considerations), security considerations were explicitly out of scope of the Provenance Working Group charter, and does not specify ways to make provenance secure.

Provenance can interact with conventional security in several ways (see [97] and works cited for further information). First, provenance might be viewed simply as data that needs to be secured, for example signed or encrypted to ensure integrity or confidentiality respectively. Second, we might view provenance as a foundation for other forms of security, for example using provenance to make judgments as to the quality or trustworthiness of some data. Finally, provenance can be viewed as a potential security risk, because blindly releasing detailed provenance may unintentionally leak confidential information.

The ability to hash and sign provenance documents is essential to determine whether documents have been tampered with,

[74] Mail 2013 Feb/0051: http://lists.w3.org/Archives/Public/public-prov-wg/2013Feb/0051.html

[75] ISSUE-425: http://www.w3.org/2011/prov/track/issues/425

and whether they have been attributed properly (See Requirement PP1). Obviously, leveraging existing standards, such as XML security[76] would be a natural approach. However, one would want a security approach to work with the idea of a conceptual model, which can be serialized in different ways. At the level of - , it would therefore become necessary to define a provenance normal form (the one discussed in - is focused on establishing logical equivalence), and ways of computing signatures, representing them, and verifying them.

If many tools and systems start using provenance, then spammers may be motivated to splatter meaningless provenance around with links to their sites. This could be extended to more malicious attempts to hinder provenance users from finding the provenance they need, or mistaking "fake" provenance for authentic. Understanding the benefits and potential security risks of provenance is an active area of research, and future versions of or standards building on may need to address security concerns more directly.

8.4. Interoperability Issue Between Serializations

While is structured according to a conceptual model and technology specific serializations (see Requirement XG2), round-trip conversions were not part of the Provenance Working Group charter. Hence, there is no requirement set on round-tripping: for instance, a translator reading an representation of , converting it to - , and back to is not required to ensure that the original representation is somehow equivalent to the final one.

Appendix A of - contains a table that cross-references the terminology used in - , - , and - . A similar table[77] makes the mapping from - to - and back fairly straightforward. However, the mapping between - in and - is more involved. During the development of , the Provenance Working Group maintained the W-ProvRDF[78] page to help keep track of the mapping between - and - / . This page was not maintained and does not reflect the final version of . Héctor Pérez-Urbina proposed a similar mapping (see M-PROV-N-RDF[79]), for a near-final version of . These may be useful as a starting point for specifying a mapping from - to and back.

Finally, for proper conversion between representations, it is likely that an agreement on basic types supported in would be required, in particular, when some serializations attempt to make the representation of some basic types such as integer more readable.

8.5. Consolidating Dictionary and Mention

Sections 7.2 and 7.7 explained how cross-bundle linking and dictionaries were moved to a note. A primary goal is to gain some experience with these constructs, ensuring they allow developers to express what they wish to represent. A secondary goal is to formalize these constructs. With cross-bundle linking, the meaning of entities (and others objects) may no longer be defined within the context of a bundle independently of other bundles. As far as dictionaries are concerned, new inferences and constraints checking should be developed.

8.6. An Expanded Vocabulary

As becomes more widely used and extended, future working groups may consider standardizing widely adopted extensions. For example, support for more comprehensive attribution or role information as it pertains to provenance may prove useful.

8.7. A Provenance API or Query language

- does not define a specific query language for provenance nor does it define an API for manipulating provenance. There are a number of query languages that have been designed for provenance [98, 99]. Furthermore, there are several APIs that have been designed to manipulate provenance[80]. While at the time of the working group many of these were in development, future working groups may find it useful to expand - to provide a common query, recording and management interface.

9. Conclusions

Some thirty years of research in provenance have culminated in a consensual view that there is a need to represent the provenance of resources and share it across the Web. With an explicit representation of provenance, the origin of such resources can be ascertained, and trust judgment can be made by their users. The design of a data model for provenance was the principal requirement set out by the charter of the Provenance Working Group. The charter suggested a list of concepts to be included in the standard, without providing definitions for them. They formed implicit requirements for the standardization activity. They constituted the Provenance Working Group's starting point, whose aim was to design a data model, as set out by its charter.

Building on a vast amount of experience with various provenance vocabularies, the Provenance Working Group participants, step by step, iteratively specified . This article captures the design decisions that influenced and the requirements that it addresses. The purpose of standardization of was not to design a comprehensive model, which was able to address all the corner cases,[81] but instead to specify what a

[76] XML Signature: http://www.w3.org/Signature/

[77] - - - : http://www.w3.org/TR/prov-xml/#prov-schema-mapping

[78] WIKI ProvRDF: http://www.w3.org/2011/prov/wiki/ProvRDF

[79] Mail PROV-N-RDF: http://lists.w3.org/Archives/Public/public-prov-comments/2013Feb/0005.html

[80] e.g. https://provenance.ecs.soton.ac.uk/store/ and https://sagebionetworks.jira.com/wiki/display/PLFM/Provenance+API

[81] The provenance work group has been amazingly effective at identifying corner cases for provenance, including famous sculptures of ice melting in the sun; legendary cakes with missing ingredients and sub-optimal oven temperatures; and customers with a red top sitting at the terrace of a nice cafe.

minimum set of constructs should be to easily address common cases. With this in mind, was designed to be extensible. The Provenance Working Group itself used the extensibility mechanism to define a few more concepts (such as dictionary, mention, and mapping to dc terms), which were regarded as useful, but not ready for Recommendation level publication. Overall, over sixty implementation reports were submitted during the implementation phase, showing a remarkable breadth of systems supporting . Finally, this article summarizes a number of outstanding issues, which may be addressed by future researchers, practitioners, and working groups.

10. Acknowledgements

The authors would like to acknowledge the contribution of all the Provenance Working Group members and the reviewers of the various specifications. This work is funded in part by the EPSRC SOCIAM (EP/J017728/1) and ORCHID (EP/I011587/1) projects, the FP7 SmartSociety (600854) project, and the ESRC eBook (ES/K007246/1) project. Additionally, this publication was supported by the Dutch national program COMMIT and the Innovative Medicines Initiative Joint Undertaking under grant agreement number 115191, resources of which are composed of financial contribution from the European Unions Seventh Framework Programme (FP7/20072013) And EFPIA companies in kind contribution.

References

[1] L. Moreau, P. Missier (eds.), K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, S. Miles, J. Myers, S. Sahoo, C. Tilmes, PROV-DM: The PROV Data Model, W3C Recommendation REC-prov-dm-20130430, World Wide Web Consortium (Oct. 2013). URL http://www.w3.org/TR/2013/REC-prov-dm-20130430/

[2] R. Bose, J. Frew, Lineage retrieval for scientific data processing: a survey, ACM Comput. Surv. 37 (1) (2005) 1–28. doi:10.1145/1057977.1057978.

[3] Y. Simmhan, B. Plale, D. Gannon, A survey of data provenance in e-science., SIGMOD Record 34 (3) (2005) 31–36. doi:10.1145/1084812.

[4] J. Cheney, L. Chiticariu, W.-C. Tan, Provenance in databases: Why, how, and where, Foundations and Trends in Databases 1 (4) (2009) 379–474. doi:10.1561/1500000006.

[5] L. Moreau, The foundations for provenance on the web, Foundations and Trends in Web Science 2 (2–3) (2010) 99–241. doi:10.1561/1800000010.

[6] L. Moreau, B. Ludaescher, I. Altintas, R. S. Barga, S. Bowers, S. Callahan, G. Chin Jr., B. Clifford, S. Cohen, S. Cohen-Boulakia, S. Davidson, E. Deelman, L. Digiampietri, I. Foster, J. Freire, J. Frew, J. Futrelle, T. Gibson, Y. Gil, C. Goble, J. Golbeck, P. Groth, D. A. Holland, S. Jiang, J. Kim, D. Koop, A. Krenek, T. McPhillips, G. Mehta, S. Miles, D. Metzger, S. Munroe, J. Myers, B. Plale, N. Podhorszki, V. Ratnakar, E. Santos, C. Scheidegger, K. Schuchardt, M. Seltzer, Y. L. Simmhan, C. Silva, P. Slaughter, E. Stephan, R. Stevens, D. Turi, H. Vo, M. Wilde, J. Zhao, Y. Zhao, The First Provenance Challenge, Concurrency and Computation: Practice and Experience 20 (5) (2008) 409–418. doi:10.1002/cpe.1233.

[7] Y. Simmhan, P. Groth, L. Moreau, Special issue: the third provenance challenge on using the open provenance model for interoperability (editorial), Future Generation Computer Systems 27 (6) (2011) 737–742. doi:10.1016/j.future.2010.11.020.

[8] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, J. Van den Bussche, The open provenance model core specification (v1.1), Future Generation Computer Systems 27 (6) (2011) 743–756. doi:10.1016/j.future.2010.07.005.

[9] Y. Gil, J. Cheney, P. Groth, O. Hartig, S. Miles, L. Moreau, P. Pinheiro da Silva, Provenance xg final report, Tech. rep., World Wide Web Consortium (2010). URL http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

[10] P. Groth, L. Moreau (eds.), PROV-Overview. An Overview of the PROV Family of Documents, W3C Working Group Note NOTE-prov-overview-20130430, World Wide Web Consortium (Apr. 2013). URL http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

[11] Y. Gil, S. M. (eds.), K. Belhajjame, H. Deus, D. Garijo, G. Klyne, P. Missier, S. Soiland-Reyes, S. Zednik, Prov model primer, W3C Working Group Note NOTE-prov-primer-20130430, World Wide Web Consortium (Apr. 2013). URL http://www.w3.org/TR/prov-primer/

[12] L. Moreau, P. Missier (eds.), J. Cheney, S. Soiland-Reyes, PROV-N: The Provenance Notation, W3C Recommendation REC-prov-n-20130430, World Wide Web Consortium (Oct. 2013). URL http://www.w3.org/TR/2013/REC-prov-n-20130430/

[13] J. Cheney, P. Missier, L. Moreau (eds.), T. D. Nies, Constraints of the PROV Data Model, W3C Recommendation REC-prov-constraints-20130430, World Wide Web Consortium (Oct. 2013). URL http://www.w3.org/TR/2013/REC-prov-constraints-20130430/

[14] T. Lebo, S. Sahoo, D. McGuinness (eds.), K. Behajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, PROV-O: The PROV Ontology, W3C Recommendation REC-prov-o-20130430, World Wide Web Consortium (Oct. 2013). URL http://www.w3.org/TR/2013/REC-prov-o-20130430/

[15] H. Hua, C. Tilmes, S. Zednik (eds.), L. Moreau, PROV-XML: The PROV XML Schema, W3C Working Group Note NOTE-prov-xml-20130430, World Wide Web Consortium (Apr. 2013). URL http://www.w3.org/TR/2013/NOTE-prov-xml-20130430/

[16] D. Garijo, K. Eckert (eds.), S. Miles, C. M. Trim, M. Panzer, Dublin Core to PROV Mapping, W3C Working Group Note NOTE-prov-dc-20130430, World Wide Web Consortium (Apr. 2013). URL http://www.w3.org/TR/2013/NOTE-prov-dc-20130430/

[17] J. Cheney, Semantics of the PROV Data Model, W3C Working Group Note NOTE-prov-sem-20130430, World Wide Web Consortium (Apr. 2013). URL http://www.w3.org/TR/2013/NOTE-prov-sem-20130430/

[18] L. Moreau, T. Lebo, Linking across provenance bundles, W3C Working Group Note NOTE-prov-sem-20130430, World Wide Web Consortium (Apr. 2013). URL http://www.w3.org/TR/2013/NOTE-prov-links-20130430/

[19] T. De Nies, S. C. (eds.), P. Missier, L. Moreau, J. Cheney, T. Lebo, S. Soiland-Reyes, Prov-dictionary: Modeling provenance for dictionary data structures, W3C Working Group Note NOTE-prov-dictionary-20130430, World Wide Web Consortium (Apr. 2013). URL http://www.w3.org/TR/2013/NOTE-prov-dictionary-20130430/

[20] G. Klyne, P. Groth (eds.), L. Moreau, O. Hartig, Y. Simmhan, J. Myers, T. Lebo, K. Belhajjame, S. Miles, PROV-AQ: Provenance Access and Query, W3C Working Group Note NOTE-prov-aq-20130430, World Wide Web Consortium (Apr. 2013). URL http://www.w3.org/TR/2013/NOTE-prov-aq-20130430/

[21] T. D. Huynh, P. Groth, S. Z. (eds.), Prov implementation report, W3C Working Group Note NOTE-prov-overview-20130430, World Wide Web Consortium (April 2013). URL http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/

[22] L. Moreau, P. Groth, Provenance: An Introduction to PROV, Morgan and Claypool, 2013. doi:10.2200/S00528ED1V01Y201308WBE007.

[23] I. Horrocks, P. Patel-Schneider, F. van Harmelen, From SHIQ and RDF to OWL: The making of a web ontology language, Web Semantics: Science, Services and Agents on the World Wide Web 1 (1). doi:10.1016/j.websem.2003.07.001.

[24] T. Baker, S. Bechhofer, A. Isaac, A. Miles, G. Schreiber, E. Summers, Key choices in the design of simple knowledge organization system (skos), Web Semantics: Science, Services and Agents on the World Wide Web 20 (0) (2013) 35 – 49. doi:10.1016/j.websem.2013.05.001.

[25] P. Groth, Y. Gil, J. Cheney, S. Miles, Requirements for provenance on

the web, International Journal of Digital Curation 7 (1). doi:10.2218/ijdc.v7i1.213.

[26] Y. R. Wang, S. E. Madnick, A polygen model for heterogeneous database systems: The source tagging perspective, in: Proceedings of the 16th International Conference on Very Large Data Bases (VLDB'90), Morgan Kaufmann, 1990, pp. 519–538. URL http://web.mit.edu/tdqm/www/tdqmpub/polygenmodelAug90.pdf

[27] P. Buneman, J. Cheney, W.-C. Tan, S. Vansummeren, Curated databases, in: PODS, 2008, pp. 1–12. doi:10.1007/978-3-642-04346-8_2.

[28] Y. Cui, J. Widom, J. L. Wiener, Tracing the lineage of view data in a warehousing environment., ACM Trans. Database Syst. 25 (2) (2000) 179–227. doi:10.1145/357775.357777.

[29] P. Buneman, S. Khanna, W. Tan, Why and where: A characterization of data provenance, in: ICDT, no. 1973 in LNCS, 2001, pp. 316–330. doi:10.1007/3-540-44503-X_20.

[30] T. J. Green, G. Karvounarakis, V. Tannen, Provenance semirings, in: PODS, 2007, pp. 31–40. doi:10.1145/1265530.1265535.

[31] F. Geerts, G. Karvounarakis, V. Christophides, I. Fundulaki, Algebraic structures for capturing the provenance of SPARQL queries, in: Joint 2013 EDBT/ICDT Conferences, ICDT '13 Proceedings, Genoa, Italy, March 18-22, 2013, 2013, pp. 153–164. doi:10.1145/2448496.2448516.

[32] P. Buneman, A. P. Chapman, J. Cheney, Provenance management in curated databases, in: Proceedings of the 2006 SIGMOD Conference on Management of Data (SIGMOD 2006), Chicago, IL, 2006, pp. 539–550. doi:10.1145/1142473.1142534.

[33] P. Buneman, S. Khanna, K. Tajima, W. Tan, Archiving scientific data, ACM Trans. Database Syst. 29 (2004) 2–42. doi:10.1145/974752.

[34] S. B. Davidson, J. Freire, Provenance and scientific workflows: challenges and opportunities, in: Proceedings of ACM Special Interest Group on Management of Data (SIGMOD 2008), 2008, pp. 1345–1350. doi:10.1145/1376616.1376772.

[35] S. Miles, P. Groth, M. Branco, L. Moreau, The requirements of recording and using provenance in e-science experiments, Journal of Grid Computing 5 (1) (2007) 1–25. doi:10.1007/s10723-006-9055-3.

[36] R. S. Barga, L. A. Digiampietri, Automatic capture and efficient storage of e-science experiment provenance, Concurrency and Computation: Practice and Experience 20 (5). doi:10.1002/cpe.1235.

[37] S. Cohen-Boulakia, O. Biton, S. Cohen, S. Davidson, Addressing the provenance challenge using zoom, Concurrency and Computation: Practice and Experience 20 (5) (2008) 497–506. doi:10.1002/cpe.1232.

[38] Y. L. Simmhan, B. Plale, D. Gannon, Karma2: Provenance management for data driven workflows, International Journal of Web Services Research 5 (2). doi:10.4018/jwsr.2008040101.

[39] S. Bowers, T. M. McPhillips, B. Ludaescher, Provenance in collection-oriented scientific workflows, Concurrency and Computation: Practice and Experience 20 (5). doi:10.1002/cpe.1226.

[40] D. Garijo, O. Corcho, Y. Gil, Detecting common scientific workflow fragments using templates and execution provenance, in: Proceedings of the Seventh International Conference on Knowledge Capture, K-CAP '13, 2013, pp. 33–40. doi:10.1145/2479832.2479848.

[41] J. Zhao, C. Goble, R. Stevens, D. Turi, Mining taverna's semantic web of provenance, Concurrency and Computation: Practice and Experience 20 (5) (2008) 463–472. doi:10.1002/cpe.1231.

[42] P. P. da Silva, L. Salayandia, A. Gates, Wdo-it! a tool for building scientific workflows from ontologies, Tech. Rep. UTEP-CS-07-XX, University of Texas at El Paso (2007). URL http://digitalcommons.utep.edu/cs_techrep/201

[43] J. Freire, C. T. Silva, S. P. Callahan, E. Santos, C. E. Scheidegger, H. T. Vo, Managing rapidly-evolving scientific workflows, in: L. Moreau, I. Foster (Eds.), Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006), Springer, 2006, pp. 10–18. doi:10.1007/11890850_2.

[44] S. Miles, P. Groth, E. Deelman, K. Vahi, G. Mehta, L. Moreau, Provenance: The bridge between experiments and data, Computing in Science and Engineering 10 (3) (2008) 38–46. doi:10.1109/MCSE.2008.82.

[45] P. Missier, S. Dey, K. Belhajjame, V. Cuevas-Vicenttín, B. Ludaescher, D-prov: Extending the prov provenance model with workflow structure, in: Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP'13), 2013, pp. 9:1–9:7. URL https://www.usenix.org/conference/tapp13/technical-sessions/presentation/missier

[46] L. Moreau, I. Foster (Eds.), Provenance and Annotation of Data — International Provenance and Annotation Workshop, IPAW 2006, Vol. 4145 of LNCS, Springer-Verlag, 2006. doi:10.1007/11890850.

[47] The First Provenance Challenge Web page (last accessed Nov 2013). URL http://twiki.ipaw.info/bin/view/Challenge/FirstProvenanceChallenge

[48] The Second Provenance Challenge Web page (last accessed Nov 2013). URL http://twiki.ipaw.info/bin/view/Challenge/SecondProvenanceChallenge

[49] Technical Summary of the Second Provenance Challenge Workshop (last accessed Nov 2013). URL http://twiki.ipaw.info/bin/view/Challenge/SecondWorkshopMinutes

[50] The Third Provenance Challenge Web page (last accessed Nov 2013). URL http://twiki.ipaw.info/bin/view/Challenge/ThirdProvenanceChallenge

[51] P. P. da Silva, D. L. McGuinness, R. Fikes, A proof markup language for semantic web services, Inf. Syst. 31 (4) (2006) 381–395. doi:10.1016/j.is.2005.02.003.

[52] D. L. McGuinness, L. Ding, P. P. da Silva, C. Chang, Pml 2: A modular explanation interlingua, in: in Proceedings of the AAAI 2007 Workshop on Explanation-aware Computing, 2007, pp. 49–55. URL https://www.aaai.org/Papers/Workshops/2007/WS-07-06/WS07-06-008.pdf

[53] S. S. Sahoo, A. Sheth, Provenir ontology: Towards a framework for escience provenance management, in: Microsoft eScience Workshop, 2009. URL http://knoesis.wright.edu/library/resource.php?id=741

[54] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg, T. Clark, The swan biomedical discourse ontology, Journal of biomedical informatics 41 (5) (2008) 739–751. doi:10.1016/j.jbi.2008.04.010.

[55] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A. J. Gray, C. Goble, T. Clark, Pav ontology: provenance, authoring and versioning, Journal of Biomedical Semantics 4 (1) (2013) 37. doi:10.1186/2041-1480-4-37.

[56] S. Sahoo, Towards desiderata for provenance ontologies in biomedicine, in: ICBO, Vol. 833, 2011. URL http://ceur-ws.org/Vol-833/paper45.pdf

[57] D. A. Bearman, R. H. Lytle, The power of the principle of provenance, Archivaria 1 (21). URL http://journals.sfu.ca/archivar/index.php/archivaria/article/viewArticle/11231

[58] O. Hartig, J. Zhao, Publishing and consuming provenance metadata on the web of linked data, in: D. L. McGuinness, J. Michaelis, L. Moreau (Eds.), IPAW, Vol. 6378 of LNCS, Springer-Verlag, 2010, pp. 78–90. doi:10.1007/978-3-642-17819-1_10.

[59] J. Zhao, Guide to the Open Provenance Model Vocabulary (2010). URL http://open-biomed.sourceforge.net/opmv/opmv-guide.html

[60] K. Alexander, M. Hausenblas, Describing linked datasets — on the design and usage of void, the vocabulary of interlinked datasets, in: In Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09), 2009. URL http://richard.cyganiak.de/2008/papers/void-ldow2009.pdf

[61] M. Schmachtenberg, C. Bizer, H. Paulheim, Adoption of the linked data best practices in different topical domains, in: The Semantic Web (ISWC'14), Vol. 8796 of LNCS, Springer-Verlag, 2014, pp. 245–260. doi:10.1007/978-3-319-11964-9_16.

[62] D. Wood, M. Lanthaler, R. Cyganiak, RDF 1.1 concepts and abstract syntax, W3C recommendation, W3C (Feb. 2014). URL http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

[63] J. J. Carroll, C. Bizer, P. Hayes, P. Stickler, Named graphs, provenance and trust, in: Proceedings of the 14th International Conference on World Wide Web (WWW'05), ACM, 2005, pp. 613–622. doi:10.1145/1060745.1060835.

[64] O. Hartig, Provenance information in the web of data, in: C. Bizer, T. Heath, T. Berners-Lee, K. Idehen (Eds.), LDOW, Vol. 538 of CEUR Workshop Proceedings, CEUR-WS.org, 2009. URL http://ceur-ws.org/Vol-538/ldow2009_paper18.pdf

[65] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, H. Shankar, Memento: Time travel for the web, arXiv

preprint. URL http://arxiv.org/abs/0911.1112

[66] A. Shaon, S. Callaghan, B. Lawrence, B. Matthews, T. Osborn, C. Harpham, A. Woolf, Opening up climate research: A linked data approach to publishing data provenance, International Journal of Digital Curation 7 (1) (2012) 163–173. doi:10.2218/ijdc.v7i1.223.

[67] K. Eckert, Provenance and Annotations for Linked Data, in: DCMI International Conference on Dublin Core and Metadata Applications (DC-2013), 2013. URL http://dcevents.dublincore.org/IntConf/dc-2013/paper/download/154/71

[68] G. Flouris, I. Fundulaki, P. Pediaditis, Y. Theoharis, V. Christophides, Coloring RDF triples to capture provenance, in: Proceedings of the 8th International Semantic Web Conference (ISWC'09), Springer-Verlag, 2009, pp. 196–212. doi:10.1007/978-3-642-04930-9_13.

[69] G. Karvounarakis, I. Fundulaki, V. Christophides, Provenance for linked data, in: In Search of Elegance in the Theory and Practice of Computation, Vol. 8000 of LNCS, Springer-Verlag, 2013, pp. 366–381. doi:10.1007/978-3-642-41660-6_19.

[70] H. Halpin, J. Cheney, Dynamic provenance for SPARQL updates, in: The Semantic Web (ISWC'14), Vol. 8796 of LNCS, Springer-Verlag, 2014, pp. 425–440. doi:10.1007/978-3-319-11964-9_27.

[71] M. Wylot, P. Cudre-Mauroux, P. Groth, TripleProv: Efficient processing of lineage queries in a native RDF store, in: 23rd International Conference on World Wide Web (WWW'14), ACM, New York, NY, USA, 2014, pp. 455–466. doi:10.1145/2566486.2568014.

[72] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space, Morgan and Claypool, 2011. doi:10.2200/S00334ED1V01Y201102WBE001.

[73] A. Schultz, A. Matteini, R. Isele, P. N. Mendes, C. Bizer, C. Becker, LDIF - A Framework for Large-Scale Linked Data Integration, in: 21st International World Wide Web Conference (WWW2012), Developers Track, 2012. URL http://www2012.org/proceedings/nocompanion/DevTrack_017.pdf

[74] J. Domingue, D. Fensel, J. A. Hendler, R. Studer, Perspectives workshop: Semantic web reflections and future directions, in: Perspectives Workshop: Semantic Web Reflections and Future Directions, no. 09271 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2010. URL http://drops.dagstuhl.de/opus/volltexte/2010/2533

[75] S. Sahoo, P. Groth, O. Hartig, S. Miles, S. Coppens, J. Myers, Y. Gil, L. Moreau, M. Panzer, D. Garijo, Provenance vocabulary mappings, Tech. rep., World Wide Web Consortium (2010). URL http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings

[76] T. Lebo, Review of prov-xg's Provenance Vocabulary Mappings (last accessed Feb 2014). URL http://inference-web.org/wiki/Review_of_prov-xg's_Provenance_Vocabulary_Mappings

[77] T. D. Huynh, M. Jewell, A. S. Keshavarz, D. Michaelides, H. Yang, L. Moreau, The PROV-JSON Serialization. A JSON Representation for the PROV Data Model, Tech. rep., University of Southampton (2013). URL http://provenance.ecs.soton.ac.uk/prov-json/

[78] S. Davidson, B. Ludaescher, T. McPhillips, J. Freire, Provenance in scientific workflow systems, Bulletin of the Technical Committee on Data Engineering 30 (4) (2007) 44–50. URL http://sites.computer.org/debull/A07dec/susan.pdf

[79] T. McPhillips, S. Bowers, B. Ludaescher, Collection-oriented scientific workflows for integrating and analyzing biological data, in: In 3rd Intl. Workshop On Data Integration in the Life Sciences (DILS), LNCS, Springer-Verlag, 2006, pp. 248–263. doi:10.1007/11799511_23.

[80] P. Groth, S. Miles, P. Missier, L. Moreau, A proposal for handling collections in the open provenance model (Jun. 2009). URL http://mailman.ecs.soton.ac.uk/pipermail/provenance-challenge-ipaw-info/2009-June/000120.html

[81] J. Davies, D. M. Germán, M. W. Godfrey, A. Hindle, Software bertillonage - determining the provenance of software development artifacts, Empirical Software Engineering 18 (6) (2013) 1195–1237. doi:10.1007/s10664-012-9199-7.

[82] J. Gray, L. Bounegru, L. Chambers (Eds.), Data Journalism Handbook 1.0 BETA, O'Reilly Media, 2012. URL http://datajournalismhandbook.org/1.0/en/

[83] M. Wooldridge, N. R. Jennings, Intelligent Agents: Theory and Practice, Knowledge Engineering Review 10 (2). doi:10.1017/S0269888900008122.

[84] L. Moreau, Provenance-based reproducibility in the semantic web, Web Semantics: Science, Services and Agents on the World Wide Web 9 (2011) 202–221. doi:10.1016/j.websem.2011.03.001.

[85] N. Kwasnikowska, L. Moreau, J. Van den Bussche, A formal account of the open provenance model, ACM Transactions on the Web. URL http://eprints.soton.ac.uk/id/eprint/374183

[86] W. E. Johnson, Logic: Part iii (1924). URL http://www.ditext.com/johnson/intro-3.html

[87] Provenance Interchange Working Group Charter (last accessed Nov 2013). URL http://www.w3.org/2011/01/prov-wg-charter

[88] M. Minsky, Logical versus analogical or symbolic versus connectionist or neat versus scruffy, AI Mag. 12 (2) (1991) 34–51. doi:10.1609/aimag.v12i2.894.

[89] S. Abiteboul, R. Hull, V. Vianu, Foundations of Databases, Addison Wesley, 1995. URL http://webdam.inria.fr/Alice/

[90] R. Fagin, P. G. Kolaitis, R. J. Miller, L. Popa, Data exchange: semantics and query answering, Theor. Comput. Sci. 336 (1) (2005) 89–124. doi:10.1016/j.tcs.2004.10.033.

[91] R. van der Meyden, Logical approaches to incomplete information: a survey, in: J. Chomicki, G. Saake (Eds.), Logics for databases and information systems, Kluwer Academic Publishers, Norwell, MA, USA, 1998, pp. 307–356. doi:10.1007/978-1-4615-5643-5_10.

[92] J. Cheney, A. Finkelstein, B. Ludäscher, S. Vansummeren, Principles of provenance (Dagstuhl Seminar 12091), Dagstuhl Reports 2 (2) (2012) 84–113. doi:10.4230/DagRep.2.2.84.

[93] B. Motik, B. C. Grau, I. Horrocks, Z. Wu, A. Fokoue, C. Lutz, OWL 2 Web Ontology Language – Profiles, Tech. rep., W3C (2009). URL http://www.w3.org/TR/2009/REC-owl2-profiles-20091027/

[94] L. Dodds, I. Davis, Linked Data Patterns: A pattern catalogue for modelling, publishing, and consuming Linked Data, Published online (2012). URL http://patterns.dataincubator.org/book/

[95] F. Manola, E. Miller (Eds.), RDF Primer, W3C Recommendation, World Wide Web Consortium, 2004. URL http://www.w3.org/TR/rdf-primer/

[96] J. Cheney, Causality and the semantics of provenance, in: DCM, 2010, pp. 63–74. doi:10.4204/EPTCS.26.6.

[97] J. Cheney, A formal framework for provenance security, in: Proceedings of the 24th IEEE Computer Security Foundations Symposium (CSF), IEEE, 2011, pp. 281–293. doi:10.1109/CSF.2011.26.

[98] G. Karvounarakis, Z. G. Ives, V. Tannen, Querying data provenance, in: International Conference on Management of Data (SIGMOD'10), 2010, pp. 951–962. doi:10.1145/1807167.1807269.

[99] J. Zhao, Y. Simmhan, K. Gomadam, V. K. Prasanna, Querying provenance information in distributed environments, IJ Comput. Appl. 18 (3) (2011) 196–215. URL http://ceng.usc.edu/~simmhan/pubs/zhao-ijca-2011.pdf